167400-83-T

# REDUCED TOLERANCE IMAGING I

J.R. FIENUP, J.N. CEDERQUIST, T.R. CRIMMINS,
R.G. PAXMAN, and D.L. NEUHOFF
**ENVIRONMENTAL RESEARCH INSTITUTE OF MICHIGAN**
P.O. Box 8618, Ann Arbor, MI 48107

JULY 1986

Technical Report for Period
15 October 1984-14 October 1985







Avionics Laboratory
Air Force Wright Aeronautical Laboratories
Air Force Systems Command
Wright-Patterson Air Force Base, Ohio 45433

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE

AD-A170 732

## REPORT DOCUMENTATION PAGE

| | |
|---|---|
| 1a. REPORT SECURITY CLASSIFICATION<br>Unclassified | 1b. RESTRICTIVE MARKINGS<br>(none) |
| 2a. SECURITY CLASSIFICATION AUTHORITY | 3. DISTRIBUTION/AVAILABILITY OF REPORT<br>Approved for public release; distribution unlimited. |
| 2b. DECLASSIFICATION/DOWNGRADING SCHEDULE | |
| 4. PERFORMING ORGANIZATION REPORT NUMBER(S)<br>167400-83-T | 5. MONITORING ORGANIZATION REPORT NUMBER(S) |
| 6a. NAME OF PERFORMING ORGANIZATION<br>Environmental Research Institute of Michigan<br>6b. OFFICE SYMBOL (If applicable) | 7a. NAME OF MONITORING ORGANIZATION<br>Air Force Wright Aeronautical Laboratories/ AARI |
| 6c. ADDRESS (City, State and ZIP Code)<br>P.O. Box 8618<br>Ann Arbor, MI 48107 | 7b. ADDRESS (City, State and ZIP Code)<br>Wright Patterson AFB, Ohio 45433 |
| 8a. NAME OF FUNDING/SPONSORING ORGANIZATION<br>Defense Advanced Research Projects Agency<br>8b. OFFICE SYMBOL (If applicable)<br>TTO | 9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER<br>F33615-83-C-1046, DARPA Order 5205 |
| 8c. ADDRESS (City, State and ZIP Code)<br>1400 Wilson Blvd.<br>Arlington, VA 22209 | 10. SOURCE OF FUNDING NOS.<br>PROGRAM ELEMENT NO. / PROJECT NO. / TASK NO. / WORK UNIT NO |
| 11. TITLE (Include Security Classification)<br>Reduced Tolerance Imaging I | |
| 12. PERSONAL AUTHOR(S)<br>J.R. Fienup, J.N. Cederquist, T.R. Crimmins, R.G. Paxman, and D.L. Neuhoff | |
| 13a. TYPE OF REPORT<br>Interim Technical<br>13b. TIME COVERED<br>FROM 10/15/84 TO 10/14/85 | 14. DATE OF REPORT (Yr., Mo., Day)<br>1986, July<br>15. PAGE COUNT<br>ix + 184 |
| 16. SUPPLEMENTARY NOTATION<br>Project Monitored by Lt. Robert Fetner, Mr. William Martin and Lt. Michael Roggemann | |

| 17 COSATI CODES | | |
|---|---|---|
| FIELD | GROUP | SUB. GR. |
| 20 | 06 | |
| 20 | 14 | |

18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
Phase retrieval
Image reconstruction

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

The reduced tolerance imaging concept is to use imaging system hardware of reduced complexity to make phase-error-degraded measurements (or to lose phase information altogether) and then reconstruct diffraction-limited imagery in a post-detection processing stage using a phase retrieval algorithm. In the first year of a two-year effort several advances were made toward this end. An estimation theoretic (Cramer-Rao) lower bound on the error of estimating a coherent image from far-field (Fourier) intensity (squared modulus) measurements was derived for the case of Gaussian detector noise. Uniqueness of reconstruction from Fourier modulus assuming a priori known support was proven for a particular class of objects -- sampled objects whose support (the area outside of which it is zero) has a convex hull with no parallel sides. A closed-form recursive reconstruction algorithm was developed for reconstructing such objects via their autocorrelation functions. Simulations showed the closed-form solution to be sensitive to noise compared with iterative Fourier transform algorithms. The uniqueness proof was useful for predicting support constraints

| | |
|---|---|
| 20. DISTRIBUTION/AVAILABILITY OF ABSTRACT<br>UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT ☐ DTIC USERS ☐ | 21. ABSTRACT SECURITY CLASSIFICATION<br>Unclassified |
| 22a. NAME OF RESPONSIBLE INDIVIDUAL<br>Mr. William Martin | 22b. TELEPHONE NUMBER (Include Area Code)<br>(513)255-6361<br>22c. OFFICE SYMBOL<br>AARI |

DD FORM 1473, 83 APR EDITION OF 1 JAN 73 IS OBSOLETE.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE

19. for which iterative reconstruction is facilitated. Several potential constraints for use in reconstruction algorithms were examined briefly, but support and non-negativity are the only two constraints that have been extensively exploited. Convergence problems when the support constraint, imposed on the world by active illumination, has tapered edges were ameliorated by a modification to the iterative transform algorithm using an "expanding mask." Alternative reconstruction algorithms were studied, including various gradient search algorithms (for which analytic expressions for the gradient of the error metric were derived) and a modelling approach, but they have not yet been developed to the point where they outperform the iterative transform algorithm. Laboratory experiments have been planned, starting with an active laser illumination of the target with a known illumination pattern and Fourier intensity measurements. Laboratory experimental set up was begun.

# PREFACE

The work reported here was performed in the Optical Science Laboratory of the Advanced Concepts Division, Environmental Research Institute of Michigan (ERIM). The work was sponsored by the Defense Advanced Projects Agency (DARPA) through the Air Force Wright Aeronautical Laboratories (AFWAL) under Contract F33615-83-C-1046, DARPA Order 5205. At AFWAL/AARI, the project monitor was first Lt. Michael Roggemann, and then the project engineer was Lt. Robert Fetner and the program manager was Mr. William Martin.

This interim technical report covers work performed from 15 October 1984 to 14 October 1985. The principal investigator at ERIM was James R. Fienup. Major contributors to this work were Jack C. Cederquist, Thomas R. Crimmins, John D. Gorman and Richard G. Paxman. Additional contributors were Carl C. Aleksoff, Greg A. Dale, Jack M. Keoshian, Timothy Klepaczyk, William H. Licata, Ivan J. LaHaie, David L. Neuhoff, Stanley R. Robinson, Nicola S. Subotic, Anthony M. Tai and Christopher C. Wackerman.

CONTENTS

LIST OF FIGURES

## LIST OF TABLES

# 1
# INTRODUCTION AND OVERVIEW

## 1.1 BACKGROUND

In many imaging scenarios that require fine resolution at long ranges, phase errors limit the achievable resolution and prevent diffraction-limited imaging. The phase errors may arise from a variety of sources, including atmospheric turbulence, misaligned or aberrated optics, motion compensation errors, local oscillator errors, and waveform generator errors. The conventional approach for obtaining diffraction-limited imagery is to build increasingly more complex sensor hardware having tight tolerances on its various components to achieve the desired phase stability.

An alternative approach is to build hardware having reduced tolerances on its phase stability, and correct for the phase errors by employing a phase retrieval algorithm in a post-processing step. In some instances a sensor can be used that is capable of measuring intensity only and does not measure the phase. Then a phase retrieval algorithm is used to retrieve the lost phase. This is what we refer to as Reduced Tolerance Imaging (RTI). Using this approach one can potentially achieve diffraction-limited imagery using a sensor system that is less complex, cheaper, lighter weight and less bulky.

In order for a phase retrieval algorithm to work, it is necessary to have some form of a priori information about, or constraints on, the image. Examples of such constraints that have been useful in the past are nonnegativity (applicable to incoherent imaging) and knowledge of the object's support (knowing its width or shape, which is available for objects on dark backgrounds or if one controls the pattern of radiation that illuminates the object).

Several important issues must be addressed to make the RTI concept feasible. Constraints must be found that are powerful enough to ensure that the retrieved phase and the reconstructed image are uniquely related to the measured data. The relationship between the reconstructed image and the measured data must be robust enough that it is not overly sensitive to noise or other imperfections in the data or constraints. Reconstruction algorithms must be found that converge reliably to a solution with a reasonable amount of computation and in the presence of realistic amounts of noise.

This report describes the results of the first year of a two-year program to develop the Reduced Tolerance Imaging concept.

## 1.2 OVERVIEW OF ACCOMPLISHMENTS TO DATE

In this section the principal results of the first year of the RTI program will be briefly summarized. They are reported in detail in the sections and appendices that follow.

One would like to know how well one could ever hope to reconstruct an image from given data and constraints. Then one would know whether current reconstruction algorithms are good enough or further development is needed. One would also be able to evaluate and compare various measurement schemes without having to develop reconstruction algorithms for each. This can be done using estimation-theoretic lower bounds on the reconstruction errors. The Cramer-Rao lower bound was derived for the case of far-field intensity measurements with additive Gaussian noise. The lower bound was computed and compared with actual errors experienced in imagery reconstructed from simulated data. These results demonstrate the usefulness of estimation theory for predicting system performance. Section 2 and Appendix D describe these results.

For discrete, or sampled, objects of a certain type a closed-form recursive reconstruction algorithm has been developed. It reconstructs an image from the autocorrelation function which is gotten by inverse Fourier transforming the measured Fourier intensity data. Although the

closed-form reconstruction algorithm has questionable usefulness because it is sensitive to noise, it has provided valuable insights into the reconstruction problem. It constitutes a uniqueness proof for the class of objects for which it is applicable and suggests illumination pattern shapes that are advantageous. These results are described in Section 3 and Appendices A, B and C.

Since image reconstruction with degraded Fourier phase or no Fourier phase requires a priori constraints on the object, it is imperative that object constraints that are sufficiently powerful and robust be found. The vast majority of the work to date has concentrated on two constraints: support, or shape (which occurs naturally for imaging satellites and may be forced by an illumination pattern) and nonnegativity (valid for most passive incoherent imaging problems). Issues relating to these and other potential constraints are discussed in Section 4.

When a support constraint is imposed by using an active illumination pattern at the target to achieve the desired known shape, the principal problem is diffraction effects at the edges of the illumination pattern. This makes the illumination pattern fall off slowly and smoothly, i.e., is tapered, rather than falling off abruptly as would be preferred. It has been found that reconstruction is much easier when there is little or no tapering of the illumination pattern. Previous versions of the iterative reconstruction algorithm were unsuccessful in reconstructing complex-valued images when large amounts of taper was present. Improved versions of the algorithm, employing an "expanding mask," were developed, and this resulted in a greatly improved result. It consists of using an unrealistically small support constraint for early iterations, which forces the energy of the image to be better centered within the true support constraint, and using progressively larger support constraints for later iterations. Section 5 describes the effects of different types of illumination patterns, describes the improved algorithm employing the expanding mask, and shows

experimental reconstruction results.

The iterative algorithm described in Section 5 is one of several possible approaches to solving the phase retrieval problem. Improved algorithms are sought which are faster and more robust. One family of alternative algorithms are the gradient search algorithms. They consist of defining a merit function, computing the gradient of the merit function as a function of a parameter space, and searching in the parameter space for a minimum of the merit function in the direction of the negative of the gradient (the global minimum of the merit function defines the solution, the reconstructed image). Merit functions that were examined include the amount by which the modulus of the Fourier transform of an object estimate differs from the measured Fourier modulus data and the amount by which an output image violates the object-domain constraints. Parameter spaces that were investigated include the space of object estimates and the space of Fourier phase estimates. Closed-form expressions for the gradients were derived, and the entire gradients can be efficiently computed using a small number of fast Fourier transforms. Gradient search algorithms were tested on the computer with mixed results to date, but they show promise and will be developed further. These results are described in Section 6 and Appendices E, F and G.

Another approach to solving the phase retrieval problem is a modeling approach. The complex Fourier transform or pieces of it are modeled by a parameterized function. The measured Fourier modulus is least-squares fitted to the modulus of the model to determine the unknown parameters. Then the parameters are inserted into the complex model which is evaluated to determine the phase. Attempts to make the modeling approach work were unsuccessful. It is likely that the models used were not appropriate to the complex Fourier transforms of interest. Better models would be needed before further work along these lines should be pursued. This work is discussed in Section 7.

The vast majority of the phase retrieval work prior to the current effort revolved around analysis and computer simulations. Since the computer simulations implicitly assume a discrete model for the object, there is a danger that the real, continuous world might behave differently. For this and other reasons it is very important tc demonstrate feasibility on real data collected in the laboratory that allows us to include the important real-world effects on the data. At least two experiments will be performed: an active, coherent experiment and a passive incoherent experiment. The active coherent experiment is being set up in the laboratory. It includes the illumination of the target with a laser beam pattern having the desired illumination shape and controlled amounts of edge tapering. A lens forms the far-field (Fourier transform) at a detector plane. A second channel including imaging optics will be used to form a "ground truth" image. Section 8 describes the active coherent experiment being set up and mentions plans for the passive incoherent experiment.

# 2
# INFORMATION THEORETIC LOWER BOUND FOR PHASE RETRIEVAL

## 2.1 INTRODUCTION

In phase retrieval problems, it is desired to estimate the phase of the Fourier transform of an object given measurements of the magnitude (i.e., the modulus or the square root of the intensity) of the Fourier transform. This is equivalent to estimating the object itself because of the Fourier transform relationship. Several iterative Fourier transform algorithms have had great success in making such object estimates from Fourier magnitude data and object constraint information [2.1, 2.2]. However, other than through empirical results [2.3], it has not been known how the error in the object estimate depends on measurement noise, constraint information, and other parameters describing the problem.

Results in estimation theory include a number of methods whereby lower bounds on the mean-squared error of the object estimate may be calculated. These methods use knowledge of the measurement procedure, the statistics of the object, and the statistics of the noise process to compute an error lower bound. An important feature is that these methods do not require specification of the algorithm used to compute the object estimate from the measured data. The lower bound, then, is independent of the algorithm and therefore indicative of the best possible estimation performance given the chosen measurements and the underlying statistics.

The Cramer-Rao lower bound is the most often used lower bound because it is usually the least difficult to compute. It has been used in a large number of single and multiple parameter and time-varying waveform estimation problems with great success [2.4]. Algorithms exist

which produce estimates that achieve the Cramer-Rao bound in problems in which the measurements are linearly related to the parameters to be estimated, the noise is additive, and the statistics are Gaussian. In nonlinear problems (of which phase retrieval will be an example), the lower bound can usually be achieved only at high signal-to-noise ratios [2.4, 2.5]; nonetheless, the lower bound is generally regarded as an important first step in evaluating and designing measurement procedures and parameter estimation algorithms for these problems. The application of lower bounds to two-dimensional signal recovery problems described here is a recent development, and it is shown that it is again a useful tool. Appendix D gives further background material on Cramer-Rao lower bounds.

## 2.2 PHASE RETRIEVAL PROBLEM DEFINITION

From the many combinations of possible phase retrieval problems and underlying assumptions, the following specific example is chosen. It it desired to estimate a two-dimensional, complex-valued object $f_m$ from real-valued measurements $S_p$ where $m = (m_1, m_2)$; $m_1, m_2 = 0, 1, \ldots, M-1$ and $p = (p_1, p_2)$; $p_1, p_2 = 0, 1, \ldots, 2M-1$. The measurements are related to the object by

$$S_p = cI_p + N_p \tag{2-1}$$

and

$$I_p = \left| \sum_m w_m f_m \exp\left[\frac{-12\pi \langle m, p\rangle}{2M}\right]\right|^2 \tag{2-2}$$

where $I_p$ is the magnitude-squared (intensity) of the discrete Fourier transform of f, c is a proportionality constant, $N_p$ is additive noise, $\langle m, p\rangle = m_1p_1 + m_2p_2$, and summation over m implies the double summation over $m_1$ and $m_2$. Object constraint information is essential for

estimating the object. The weighting array $w_m$ is explicitly included in Eq. (2-2) to allow arbitrary support constraints to be placed on the object. For an object of M by M resolution elements, Nyquist sampling requires a measurement array of size 2M by 2M because the magnitude-squared has twice the bandwidth of the complex-valued Fourier transform. It will be convenient later to consider w, f, S, I, and N as vectors. The phase retrieval problem is to estimate the object f given the set of measurements S and knowledge of the constraint that the product $w_m f_m$ is zero wherever $w_m$ is known to be zero.

This mathematical statement can represent a number of applications in which phase retrieval problems arise. For example, consider the measurement geometry shown in Fig. 2-1. A known, complex-valued, monochromatic wavefront w(x,y) illuminates an unknown, complex-valued object f(x,y). Alternatively, for the wavefront sensing problem, an unknown monochromatic wavefront may pass through a known aperture having known complex transmittance w(x,y). The intensity I(u,v) in a measurement plane located a distance R from the object plane is:

$$I(u, v) = \frac{1}{(\lambda R)^2} \left| \iint w(x, y) f(x, y) \exp\left[\frac{-i2\pi(ux + vy)}{\lambda R}\right] dx\, dy \right|^2 \quad (2\text{-}3)$$

where $\lambda$ is the wavelength and it is assumed that R is sufficiently great that the Fraunhofer approximation can be made. A discrete set of measurements S is made with

$$S_p = \eta T \int_{\Delta A} I(u, v)\, du\, dv + N_p \quad (2\text{-}4)$$

where $\eta$ is the detector efficiency, T is the detector integration time, $\Delta A$ is the area of a detector element, $N_p$ is the detector noise, and the subscript $p = (p_1, p_2)$ indexes over the measurement plane. A phase retrieval method (e.g., an iterative Fourier transform algorithm) would be applied to the measurement set S using the object constraint provided

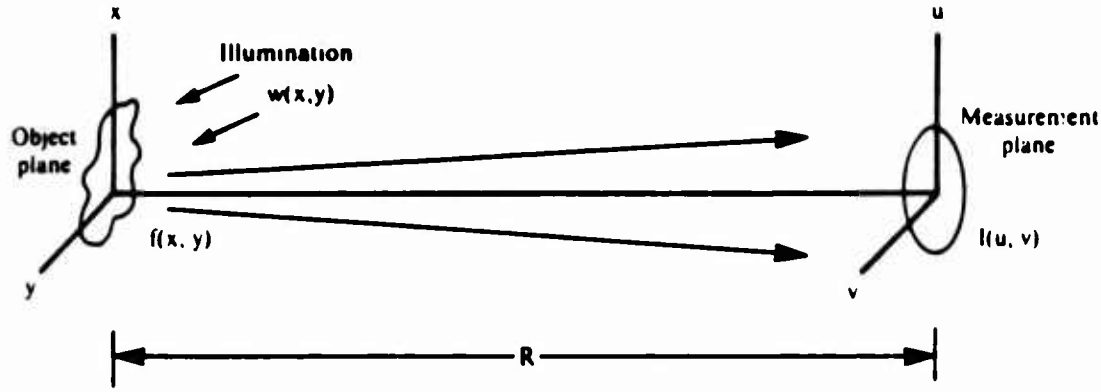



Figure 2-1. (U) Measurement Geometry for Phase Retrieval Problem

by the illumination pattern w to give an estimate of a sampled version of the object f. Conversion of Eqs. (2-3) and (2-4) into discrete form gives, for this application, a value for the constant c in Eq. (2-1) of $\eta T \Delta A (\Delta a / \lambda R)^2$ where $\Delta a$ is the square area of an object sample.

The complex-valued object f can be written in terms of its real and imaginary parts,

$$f_m = f_m^r + i f_m^i. \tag{2-5}$$

Equation (2-2) then becomes

$$I_p = \left| \sum_m w_m \left( f_m^r + i f_m^i \right) \exp \left[ \frac{-i\pi \langle m, p \rangle}{M} \right] \right|^2. \tag{2-6}$$

## 2.3 CRAMER-RAO LOWER BOUND

It can be proven that the variance of any unbiased estimate of a component of a random vector is greater than or equal to the corresponding diagonal element of the inverse of what is called the Fisher information matrix. The value of the diagonal element is the Cramer-Rao lower bound. The elements of the Fisher information matrix depend in turn upon the second partial derivatives of the joint probability distribution of the measurement vector and the vector to be estimated. This result is proven primarily by the use of the Schwarz inequality.

Application of the Cramer-Rao method for determining lower bounds on estimation errors to a specific problem must therefore begin with a determination of the statistics of the parameters to be estimated and of the noise [2.4, 2.6]. In this analysis, it is assumed that $f_m^r$, $f_m^i$, and $N_p$ are each statistically independent, zero mean, Gaussian random

variables with variances $\sigma_f^2/2$, $\sigma_f^2/2$, and $\sigma_N^2$ respectively. Note that this implies that the variance of the complex-valued $f_m$ is $\sigma_f^2$.

By the definition of conditional probability,

$$p(S, f) = p(S|f)p(f) \tag{2-7}$$

where $p(S,f)$ is the joint probability density of S and f, $p(S|f)$ is the conditional probability density of S given f, and $p(f)$ is the probability density of f. (Recall that f and S are vectors.) The assumption of Gaussian statistics gives

$$p(f) = \prod_m \frac{1}{\pi\sigma_f^2} \exp\left[\frac{-|f_m|^2}{\sigma_f^2}\right] \tag{2-8}$$

and, using Eqs. (2-1) and (2-6) which imply that $p(S|f) = p(N = S - cI)$,

$$p(S|f) = \prod_p \frac{1}{\sigma_n\sqrt{2\pi}} \exp\left[\frac{-(S_p - cI_p)^2}{2\sigma_N^2}\right]. \tag{2-9}$$

The Cramer-Rao method continues by defining the Fisher information matrix J in terms of the probability density functions. For the present problem, where it is desired to estimate the statistically independent real and imaginary parts of f, a workable notation is to partition J into four submatrices:

$$J = \left[\begin{array}{c:c} J^{rr} & J^{ri} \\ \hdashline J^{ir} & J^{ii} \end{array}\right]. \tag{2-10}$$

J is of dimension $2M^2$ by $2M^2$ (representing the $M^2$ independent $f_m^r$ plus the $M^2$ independent $f_m^i$) and each of the submatrices is of dimension $M^2$ by $M^2$. The elements of the submatrices are defined by, for example [2.4, 2.6],

$$J_{mn}^{rr} = -E\left[\frac{\partial^2 \ell n\, p(S, f)}{\partial f_m^r \partial f_n^r}\right] \tag{2-11}$$

where $E[\cdot]$ denotes expectation taken over both f and N and the partial derivative holds S constant. The other submatrices are defined by appropriate substitution of the superscripts r and i. It is assumed that these and any other required derivatives exist. This assumption is valid for the phase retrieval problem.

The Cramer-Rao method concludes by determining the inverse $J^{-1}$ of the Fisher information matrix J. The diagonal elements of $J^{-1}$ are the desired lower bounds on the mean-squared error of the object estimate $\hat{f}$. From the convention used to define J, the upper left diagonal elements of $J^{-1}$ refer to $f_m^r$ and the lower right elements to $f_m^i$. If $J^{-1}$ is similarly partitioned into four submatrices:

$$J^{-1} = \left[\begin{array}{c:c} K^{rr} & K^{ri} \\ \hdashline K^{ir} & K^{ii} \end{array}\right], \tag{2-12}$$

then the Cramer-Rao lower bound $e_{0m}^2$ on the mean-squared error, $E[|\hat{f}_m - f_m|^2]$, in the estimate $\hat{f}_m$ of object component $f_m$, is the sum of the diagonal elements for $f_m^r$ and $f_m^i$:

$$e_{om}^2 = K_{mm}^{rr} + K_{mm}^{ii}. \tag{2-13}$$

This is the quantity which the following analysis seeks. Strictly, the lower bound is only for unbiased estimates of f. It is beyond the scope of this work to determine whether particular phase retrieval algorithms give unbiased estimates.

## 2.4 LOWER BOUND FOR PHASE RETRIEVAL

Substituting Eqs. (2-8) and (2-9) into Eq. (2-11), differentiating, and discarding a term with zero expected value gives [2.4]

$$J_{mn}^{rr} = \frac{c^2}{\sigma_N^2} \sum_p E\left[\frac{\partial I_p}{\partial f_m^r} \frac{\partial I_p}{\partial f_n^r}\right] + \frac{2\delta_{mn}}{\sigma_f^2} \tag{2-14}$$

where $\delta_{mn}$ is the Kronecker delta function. Similar results hold for the other submatrices of J except that $J^{ri}$ and $J^{ir}$ have no $\delta_{mn}$ term. It is important to note that this result holds for any function I of the parameter f. It does not assume that the measurements are of the Fourier magnitude-squared.

Equation (2-6) can now be used to compute the first term on the right hand side of Eq. (2-14). First,

$$\frac{\partial I_p}{\partial f_m^r} = w_m^* \sum_j w_j (f_j^r + i f_j^i) \exp\left[\frac{-i\pi\langle j - m, p\rangle}{M}\right] + c.c. \tag{2-15}$$

Then,

$$\sum_p E\left[\frac{\partial I_p}{\partial f_m^r} \frac{\partial I_p}{\partial f_n^r}\right] =$$

$$\sum_p E\left[w_m^* w_n^* \sum_j \sum_k w_j w_k (f_j^r + i f_j^i)(f_k^r + i f_k^i) \exp\left[\frac{-i\pi\langle j + k - m - n, p\rangle}{M}\right]\right.$$

$$+ w_m^* w_n \sum_j \sum_k w_j w_k^* (f_j^r + i f_j^i)(f_k^r - i f_k^i) \exp\left[\frac{-i\pi\langle j - k - m + n, p\rangle}{M}\right]$$

$$+ \text{c.c.'s}\Bigg]. \qquad (2\text{-}16)$$

Taking the expected value gives

$$\sum_p E\left[\frac{\partial I_p}{\partial f_m^r}\frac{\partial I_p}{\partial f_n^r}\right] =$$

$$\sum_p w_m^* w_n \sum_j |w_j|^2 \sigma_f^2 \exp\left[\frac{-i\pi\langle n - m, p\rangle}{M}\right] + \text{c.c.} \qquad (2\text{-}17)$$

The summation over k is eliminated since the $f_m^r$ and $f_m^i$ are independent. The first and third terms in Eq. (2-16) are also eliminated because the $f_m^r$ and $f_m^i$ have equal variances. Finally, the summation over p gives

$$J_{mn}^{rr} = \left(\frac{8c^2\sigma_f^2 M^2 |w_m|^2}{\sigma_N^2}\sum_j |w_j|^2 + \frac{2}{\sigma_f^2}\right)\delta_{mn} \qquad (2\text{-}18)$$

because

$$\sum_p \exp\left[\frac{-i\pi\langle n - m, p\rangle}{M}\right] = 4M^2\delta_{mn}. \qquad (2\text{-}19)$$

Equation (2-18) is a general expression for one of the submatrices of the Fisher information matrix J given the assumptions above. Similar computations show that $J^{ii} = J^{rr}$ and $J^{ri} = J^{ir} = 0$. In this case, then, J is diagonal and can be analytically inverted to obtain $J^{-1}$. This is, of course, a result of the discrete Fourier transform nature of Eq. (2-2). Other phase retrieval problems may lead to nondiagonal J matrices which may be difficult or impractical to invert analytically.

Using Eqs. (2-13) and (2-18), the lower bound $e_{0m}^2$ on the mean-squared error in the estimate of $f_m$ is:

$$e_{om}^2 = \frac{\sigma_f^2}{1 + \dfrac{4c^2\sigma_f^4 M^2 |w_m|^2}{\sigma_N^2} \displaystyle\sum_j |w_j|^2} \quad . \tag{2-20}$$

It is, as stated earlier, independent of the phase retrieval algorithm used to estimate f.

The notation of Eq. (2-20) can be simplified by defining a signal-to-noise ratio:

$$SNR = \frac{\{E[cI_p]\}^2}{\sigma_N^2} \tag{2-21}$$

where, by Eq. (2-6),

$$E[cI_p] = c\sigma_f^2 \sum_j |w_j|^2. \tag{2-22}$$

Equation (2-20) then becomes

$$e_{om}^2 = \frac{\sigma_f^2}{1 + \dfrac{4\ SNR\ M^2\ |w_m|^2}{\displaystyle\sum_j |w_j|^2}} \quad . \tag{2-23}$$

As would be expected, the lower bound on the estimate reduces to the a priori variance $\sigma_f^2$ if either $f_m$ is not illuminated ($w_m = 0$) or the SNR is zero. The lower bound also approaches zero as the SNR approaches infinity.

For the case in which the magnitudes of the support constraint w are either zero or one, Eq. (2-20) predicts that, if the support constraint includes a smaller part of the M by M object array (and therefore $\sum_j |w_j|^2$ decreases), then the error lower bound increases. This is due to the loss of signal as can be seen from Eq. (2-22). On the other hand, if the SNR is held constant, then Eq. (2-23) predicts that the error bound decreases. This is equivalent to sampling at greater than the Nyquist rate in the measurement array in the Fourier domain. The well-known error decrease is known as compression gain.

It is known that current iterative phase retrieval algorithms are more successful in converging to a solution for some object support constraints than for others (e.g., for a triangularly-shaped pattern imposed by w, the algorithm more readily finds a solution than for a square pattern) [2.7]. By a solution is meant an object estimate that is as close to agreeing with the measured data and the a priori constraints as possible. In some cases, an algorithm stagnates and produces an output in poor agreement with the data and constraints; such an output should not be considered an object estimate. If there is more than one solution that closely agrees with the data and constraints, the algorithm may find a solution that is different from the true object. There is a tendency for iterative transform algorithms to find solutions more readily for cases guaranteed to have unique solutions (e.g., objects with triangular support constraints). However, when the solution is unique, it is also known that, if a solution is found (i.e.,

the algorithm does not stagnate in poor agreement with the data and constraints), then the mean-squared error is independent of the shape of the object support constraint. From either Eq. (2-20) or Eq. (2-23), it can be seen that, for a given value of $\sum_j |w_j|^2$, the lower bound $e_{0m}^2$ depends only on $|w_m|^2$ and not on the two-dimensional distribution of w (the support constraint). The Cramer-Rao lower bound is apparently a measure of the error of algorithms which have found a reasonably good estimate and is insensitive to lack of uniqueness or to algorithm-dependent problems such as stagnation. The insensitivity to uniqueness is further demonstrated by an example shown in Appendix D.

## 2.5 CONCLUSION AND SUGGESTIONS FOR FURTHER RESEARCH

In this investigation of the application of estimation theoretic lower bounds to phase retrieval and image reconstruction problems, the Cramer-Rao lower bound on the mean-squared error in the object estimate from Fourier magnitude-squared measurements, given additive noise, Gaussian statistics, and Nyquist sampling, was found. The lower bound approaches the appropriate values in the limits of high and low SNR, but does not depend on the object support constraint. Further research should investigate other measurement models (e.g., Fourier magnitude measurements), object domain constraints (e.g., nonnegativity), statistical assumptions (e.g., Poisson noise), and/or other information theoretic lower bounds to extend and refine the bounds and to attempt to show a dependence on a priori knowledge such as support constraints. Computer simulations and laboratory experiments should also be performed to allow comparison of the lower bound to the error achieved by current phase retrieval algorithms.

REFERENCES

2.1. J.R. Fienup, "Reconstruction and synthesis applications of an iterative algorithm," Proc. SPIE 373, 147-160 (1981).

2.2. J.R. Fienup, "Phase retrieval algorithms: a comparison," Appl. Opt. 21, 2758-2769 (1982).

2.3. G.B. Feldkamp and J.R. Fienup, "Noise properties of images reconstructed from Fourier modulus," Proc. SPIE 231, 84-93 (1980).

2.4. H. Van Trees, Detection, Estimation, and Modulation Theory, Part I (Wiley, New York, 1968), pp. 66-73, 79-85, 437-441.

2.5. J.N. Cederquist, S.R. Robinson, D. Kryskowski, J.R. Fienup, and C.C. Wackerman, "Cramer-Rao lower bound on wavefront sensor error," Proc. SPIE 551, 146-155 (1985).

2.6. H. Van Trees, "Bounds on the accuracy attainable in the estimation of continuous random processes," IEEE Trans. Inform. Theory IT-12, 298-305 (1966).

2.7. J.R. Fienup, "Phase retrieval from a single intensity distribution," in Optics in Modern Science and Technology (ICO-13, Sapporo, Japan, 1984), pp. 606-609.

# 3
# UNIQUE CLOSED FORM RECONSTRUCTION ALGORITHM

## 3.1 INTRODUCTION

Since the object's autocorrelation function can be computed from the modulus of its Fourier transform, reconstructing the object from its autocorrelation is equivalent to reconstructing it from the modulus of its Fourier transform. In an earlier effort, it was shown that a unique closed-form algorithm for reconstructing an object from its autocorrelation, which operated in a recursive fashion, was possible for two very special kinds of objects: those fitting within a rectangle with an additional point off one corner of the rectangle and those fitting within a triangle having nonzero corners. This earlier result has been vastly generalized to include objects having supports whose convex hulls have no parallel sides, a very large class of objects. This generalized algorithm, which includes a uniqueness proof, is described in Section 3.2 and Appendicies A, B and C.

Experimental reconstruction results obtained using the algorithm are shown in Section 3.3. Although the present form of the algorithm is very sensitive to noise, limiting its practical use, it has proven to be very valuable in that it suggests useful illumination pattern (support) constraints, as is demonstrated in Section 4.1. Another problem with this reconstruction algorithm is that it explicitly assumes a sampled object, i.e. one consisting of an array of delta functions, and it cannot in its present form be employed for real-world continuous objects. One possible way around this problem is to use the quasi-sampling method suggested in Section 3.4.

## 3.2 PHASE RETRIEVAL FOR DISCRETE FUNCTIONS WITH SUPPORT CONSTRAINTS

### 3.2.1. INTRODUCTION

The reconstruction of object functions having non-redundant spacings was discussed in [3.1]. Hayes and Quatieri [3.2] showed that the boundaries of triangular objects can be reconstructed by making use of certain spacings in the object which are non-redundant. In another direction, Bruck and Sodin [3.3] showed that the uniqueness of phase retrieval is equivalent to the irreducibility of a polynomial in two variables which is closely related to the Fourier transform (z-transform) of the object function. Fiddy, Brames and Dainty [3.4] used Eisenstien's irreducibility criterion to prove uniqueness for object functions satisfying certain support constraints and showed that Fienup's input-output iterative Fourier transform algorithm [3.5-3.7] converged faster to a better reconstruction when these constraints were satisfied. Fienup [3.8] presented a closed-form algorithm for reconstructing such object functions from their autocorrelation functions. He also presented a similar closed-form reconstruction algorithm for objects satisfying a triangular support constraint and thereby showed that such objects are uniquely defined by their autocorrelation functions among all object functions satisfying the same support constraint.

A generalization of Fienup's results to a wider class of support constraints is presented here. Also, an algorithm for generating closed-form reconstruction algorithms is described. Brames [3.9]

recently obtained a result similar to the uniqueness theorem in Section 3.2.3.

### 3.2.2. MASKS

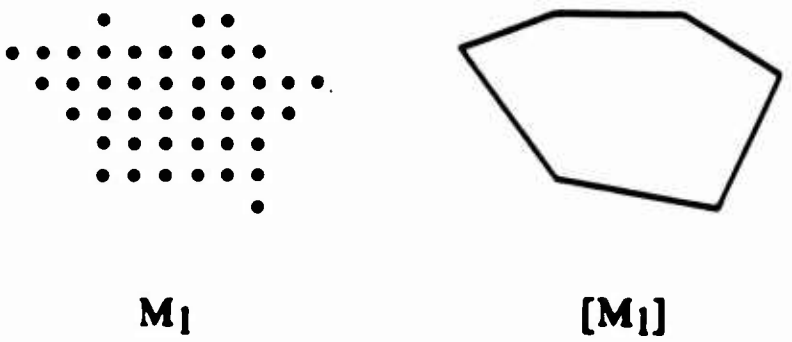
Let $\mathcal{R}^2$ denote the Euclidean plane and let $Z^2$ denote the points in $\mathcal{R}^2$ with integer coordinates. A finite subset of $Z^2$ is a <u>mask</u> if it contains at least three non-collinear points and its convex hull in $\mathcal{R}^2$ (the smallest convex set containing it) has no parallel sides. Let M be a mask and let [M] denote its convex hull in $\mathcal{R}^2$. Then [M] is a convex polygon (including its interior). See Figure 3-1. A vertex v of [M] is <u>opposite</u> a side s of [M] if the line through v and parallel to s contains no points of [M] other than v (see Figure 3-2). A vertex of [M] is a <u>reference point</u> of M if it is opposite some side of [M] (see Figure 3-3). The set of all reference points of M will be denoted by R(M).

### 3.2.3. UNIQUENESS THEOREM

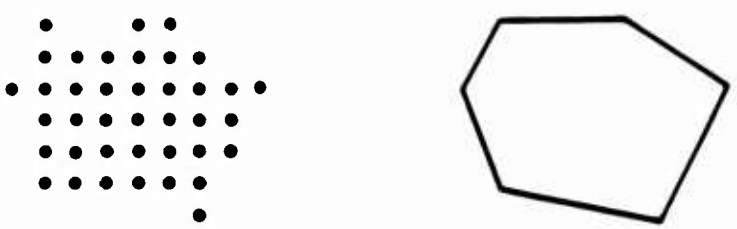
Let f be a complex-valued function on $Z^2$. The support of a function on $Z^2$ is the set of points at which the function is non-zero. Let $\mathcal{S}(f)$ denote the support of f. If $\mathcal{S}(f)$ is a finite set, the autocorrelation function of f is defined for $x \in Z^2$ by

$$r(x) = \sum_{y \in Z^2} f(y) f^*(y - x) \qquad (3\text{-}1)$$

where the * denotes complex conjugation. Let $f_1$ be another



$M_1$ $[M_1]$



$M_2$ $[M_2]$

FIGURE 3-1. $M_1$ IS A MASK. $M_2$ IS NOT A MASK SINCE $[M_2]$ HAS TWO PARALLEL SIDES.

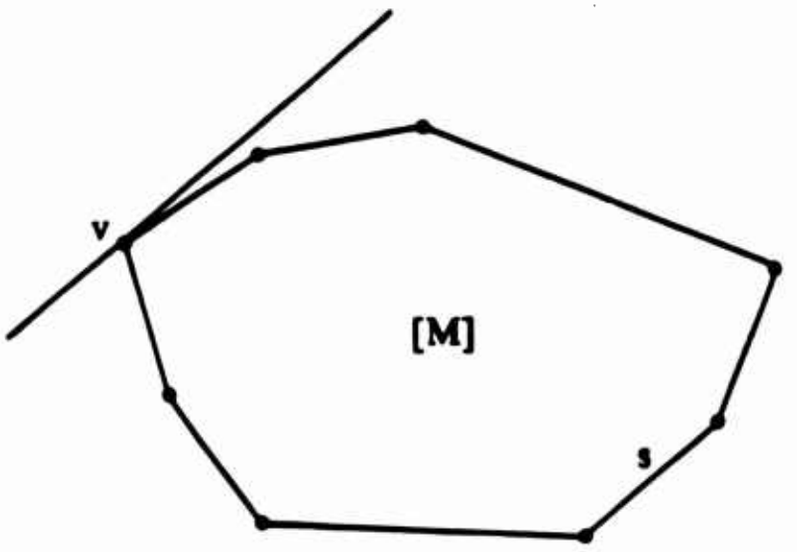



FIGURE 3-2. THE VERTEX v IS OPPOSITE THE SIDE s.

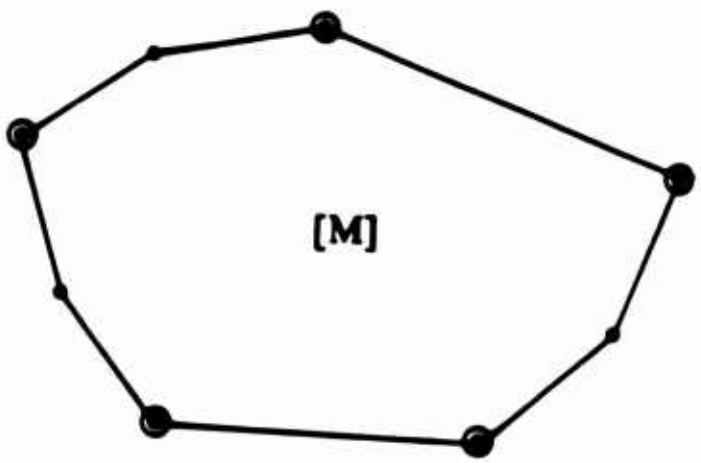



FIGURE 3-3. THE CIRCLED VERTICES OF [M] ARE THE REFERENCE POINTS OF THE MASK M.

complex-valued function on $Z^2$ with finite support $S(f_1)$ and autocorrelation function $r_1$.

We have the following uniqueness theorem.

Theorem: If M is a mask, $R(M) \subseteq S(f) \subseteq M$, $S(f_1) \subseteq M$ and $r = r_1$, then there exists a complex number $\alpha$ of modulus 1 such that $f_1 = \alpha f$.

The proof is in Appendix A.

### 3.2.4. RECONSTRUCTION ALGORITHMS

In this section closed-form algorithms for reconstructing a function from its autocorrelation function will be described.

Let S be the number of vertices of [M]. Let $v_0, \ldots, v_{S-1}$ be an ordering of the vertices going around [M] in the counter-clockwise direction and let $p_0, \ldots, p_{T-1}$ be a similar ordering of the reference points of M. By Lemma A-2 in Appendix A, R(M) contains an odd number of points so that T is odd. Let $K = (T - 1)/2$ and let $q_n = p_{(nK) \bmod T}$ for $n = 0, \ldots, T - 1$. Since K and T are relatively prime, the $q_n$ are distinct and hence run through all the points of R(M) (see Figure 3-4). By Lemma A-4 in Appendix A, $q_n$ and $q_{(n+1) \bmod T}$ have unique separation in M. That is, if x, $y \in M$ and $x - y = q_{(n+1) \bmod T} - q_n$ then $x = q_{(n+1) \bmod T}$ and $y = q_n$.

Let N be the number of points in M. A reconstruction algorithm for the mask M is an ordered pair, (q, m), where $q = (q_0, \ldots, q_{N-1})$

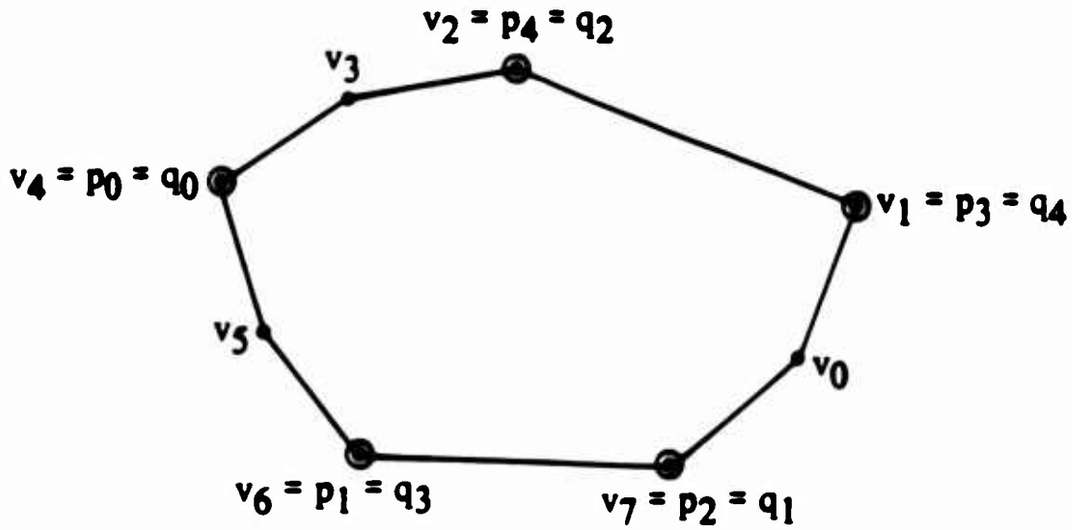



FIGURE 3-4. THE NUMBERINGS OF THE VERTICES AND REFERENCE POINTS OF A MASK. HERE S = 8, T = 5, AND K = 2.

is an ordering of the points in M and $m = (m_T, \ldots, m_{N-1})$ is a sequence of integers satisfying the following conditions. The points $q_0, \ldots, q_{T-1}$ are as described above. For $n = T, \ldots, N-1$, the integers $m_n$ satisfy the conditions $0 \leq m_n \leq T-1$, and $M \cap (M + q_n - q_{m_n}) \subseteq \{q_0, \ldots, q_n\}$ and $M \cap (M - q_n + q_{m_n}) \subseteq \{q_0, \ldots, q_{n-1}\}$. In the next section, an algorithm for generating such reconstruction algorithms will be described.

In order to justify the above definition of reconstruction algorithms it will now be shown how they can be used to reconstruct a function from its autocorrelation function.

Let f be a complex-valued function on $Z^2$ satisfying $R(M) \subseteq S(f) \subseteq M$ and let r be its autocorrelation function. Now let $x = q_{(n+1) \bmod T} - q_n$ and suppose that for some $y \in Z^2$, $f(y)f^*(y-x) \neq 0$. Then $y \in S(f) \subseteq M$ and $y - x \in S(f) \subseteq M$. Also, $y - (y - x) = x = q_{(n+1) \bmod T} - q_n$. Since $q_{(n+1) \bmod T}$ and $q_n$ have unique separation in M, it follows that $y = q_{(n+1) \bmod T}$ and $y - x = q_n$. Therefore $y = q_{(n+1) \bmod T}$ is the only $y \in Z^2$ for which $f(y)f^*(y-x) \neq 0$, hence

$$r(q_{(n+1) \bmod T} - q_n) = f(q_{(n+1) \bmod T})\, f^*(q_n), \qquad (3\text{-}2)$$

and since $R(M) \subseteq S(f)$, $r(q_{(n+1) \bmod T} - q_n) \neq 0$. It now follows from (3-2) that

$$|f(q_0)|^2 = \frac{\prod_{n=0}^{K} r(q_{2n} - q_{(2n+1) \bmod T})}{\prod_{n=0}^{K-1} r(q_{2n+2} - q_{2n+1})}. \qquad (3\text{-}3)$$

Since f is defined by r only up to multiplication by a modulus 1 complex number, we may require that $f(q_0) > 0$. Then $f(q_0)$ is equal to the positive square root of the right-hand side of (3-3). Now $f(q_n)$ can be computed for $n = 1, \ldots, T-1$ from the formula $f(q_n) = r(q_n - q_{n-1})/f^*(q_{n-1})$. It is shown in Appendix B that if (q, m) is a reconstruction algorithm then the following program will compute $f(q_n)$ for $n = T, \ldots, N-1$. Set $f(x) = 0$ for $x \in Z^2$ and $x \neq q_n$, $n = 0, \ldots, T-1$, and set $n = T-1$.

Step 1: If $n = N-1$, stop. Otherwise $n \leftarrow n+1$.

Step 2: $$f(q_n) = \left[ r(q_n - q_{m_n}) - \sum_{k=0}^{n-1} f(q_k)\, f^*(q_k - q_n + q_{m_n}) \right] / f^*(q_{m_n}).$$

Step 3: Go to Step 1.

### 3.2.5. ALGORITHM FOR GENERATING RECONSTRUCTION ALGORITHMS

It will be assumed that we are given a sequence of all vertices $v_0, \ldots, v_{S-1}$ of [M] where M is a mask and the sequence is ordered in the counter-clockwise direction around [M].

For $n = 0, \ldots, S-1$, let $s_n$ be the side of [M] with endpoints $v_n$ and $v_{(n+1) \bmod S}$. Let U be the linear operator on $\mathcal{R}^2$ which rotates each vector in $\mathcal{R}^2$ 90° counter-clockwise.

First, the reference points $p_0, \ldots, p_{T-1}$ must be found. Note that every side of [M] has a vertex opposite it which is therefore a reference point. Of course, several sides may have the same vertex opposite them. Let $w_n = v_{(n+1) \bmod S} - v_n$ for $n = 0, \ldots, S-1$.

A vertex $v_m$ is opposite a side $s_n$ if and only if $\langle v_m, Uw_n \rangle \geq \langle v_k, Uw_n \rangle$ for $k = 0, \ldots, S-1$, where $\langle , \rangle$ denotes the inner product on $\mathcal{R}^2$ (see Figure 3-5). Thus, by taking each side in the order $s_0, \ldots, s_{S-1}$, all the reference points of M will be found, and if they are numbered in the order in which they are found, $p_0, \ldots, p_{T-1}$, then the ordering will be in the counter-clockwise direction around [M].

As mentioned above T is odd. Let $K = (T-1)/2$ and $q_n = p_{(nK) \bmod T}$, $n = 0, \ldots, T-1$. Since each $q_n$ is a reference point and therefore is a vertex of [M], there exists an integer $k_n$ such that $0 \leq k_n \leq S-1$ and $q_n = v_{k_n}$. For $n = 0, \ldots, T-1$, define

$$y_n = Uw_{(k_n-1) \bmod S} - Uw_{k_{(n+1) \bmod T}}. \tag{3-4}$$

Then by Lemma A-3 in Appendix A, for $x \in M$, $x \neq q_n$ and $x \neq q_{(n+1) \bmod T}$, $\langle q_n, y_n \rangle < \langle x, y_n \rangle < \langle q_{(n+1) \bmod T}, y_n \rangle$. This is equivalent to saying all points in M excluding $q_n$ and $q_{(n+1) \bmod T}$ lie strictly between lines pendicular to $y_n$ and passing through $q_n$ and $q_{(n+1) \bmod T}$. See Figure 3-6. (The uniqueness of separation of $q_n$ and $q_{(n+1) \bmod T}$ mentioned in Section 3.2.4 follows from this double inequality.)

Now let $\alpha_n = q_n + q_{(n+1) \bmod T}$ and let $\beta_n = \alpha_n/2$ for $n = 0, \ldots, T-1$. Then $\beta_n$ is the midpoint of the line segment joining $q_n$ and $q_{(n+1) \bmod T}$. Let $D = M \setminus R(M)$ (set difference) and let $\phi$ be the characteristic function of D as a subset of $Z^2$. This is, $\phi$ is the function on $Z^2$ which is 1 on D and 0 outside D.

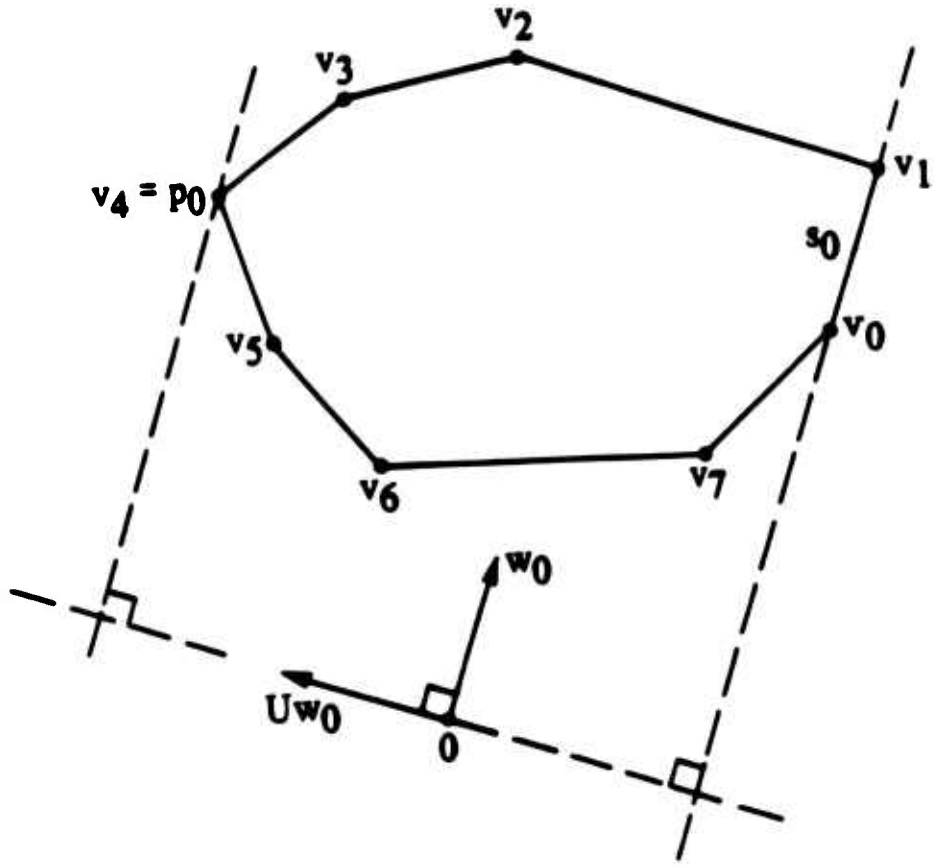



FIGURE 3-5. AN ILLUSTRATION OF THE VECTORS $w_n$ AND $Uw_n$. HERE $n = 0$ AND THE ORIGIN IN $\mathcal{R}^2$ IS DENOTED BY "0".

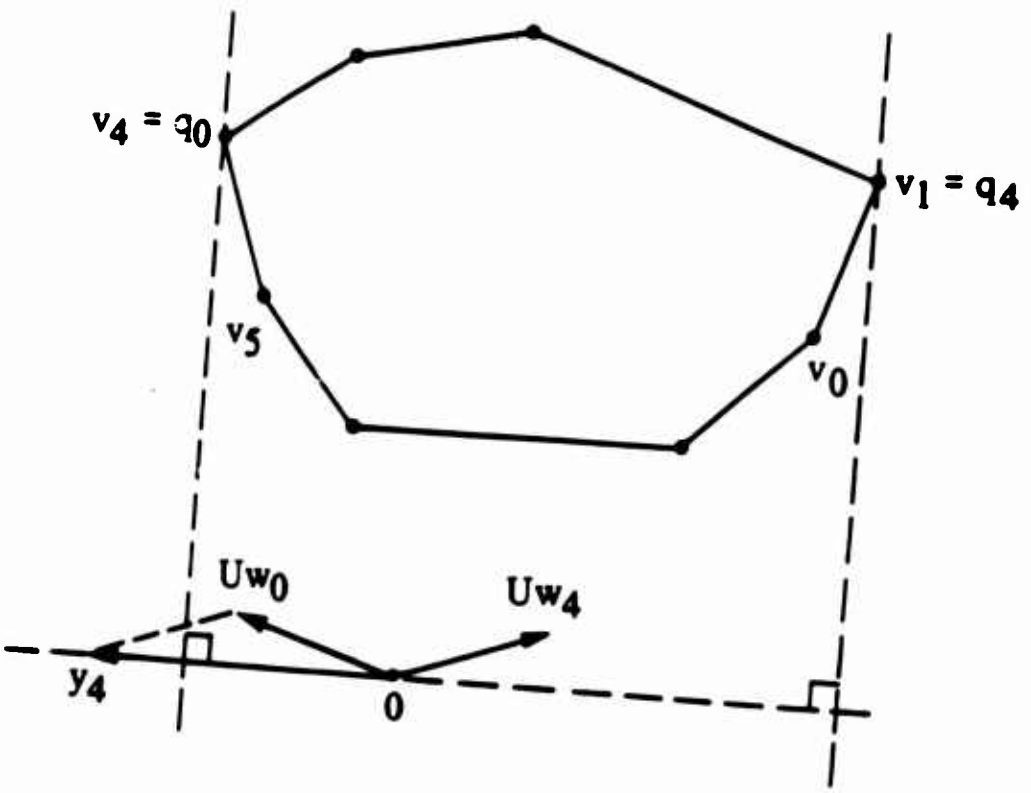



FIGURE 3-6. AN ILLUSTRATION OF THE VECTORS $y_n$. HERE S = 8, T = 5, n = 4, $k_4$ = 1, ($k_4$ - 1)mod S = 0, (n + 1)mod T = 0 AND $k_0$ = 4.

For $x \in D$ and $0 \leq n \leq T - 1$, let $h_n(x) = \langle x - \beta_n, y_n \rangle$. Then $|h_n(x)|/||y_n||$ is the distance from x to the line through $\beta_n$ and perpendicular to $y_n$, where $||y_n||$ denotes the Euclidean norm of $y_n$ (see Figure 3-7). The set D contains N - T points and we define T orderings of the points in D, $D = \{ d_{n,0}, \ldots, d_{n,N-T-1} \}$, $n = 0, \ldots, T - 1$, satisfying $|h_n(d_{n,k})| \geq |h_n(d_{n,k+1})|$ for $k = 0, \ldots, N - T - 2$. The following program generates sequences $q_T, \ldots, q_{N-1}$ and $m_T, \ldots, m_{N-1}$.

Set $n = T - 1$ and $k = 0$ and enter the following loop.

Step 1: If $n = N - 1$, stop. Otherwise define

$$b = \min \{ j : 0 \leq j \leq N - T - 1 \text{ and } \phi(d_{k,j}) = 1 \}.$$

Step 2: If $\phi(a_k - d_{k,b}) = 1$, go to Step 7.

Step 3: $n \leftarrow n + 1$.

Step 4: Define $q_n = d_{k,b}$.

Step 5: If $h_k(q_n) \geq 0$, define $m_n = k$. Otherwise define

$$m_n = (k + 1) \bmod T.$$

Step 6: $\phi(q_n) \leftarrow 0$.

Step 7: $k \leftarrow (k + 1) \bmod T$ and go to Step 1.

It is shown in Appendix C that the loop is not infinite and if $q = (q_0, \ldots, q_{N-1})$ and $m = (m_T, \ldots, m_{N-1})$ then $(q, m)$ is a reconstruction algorithm.

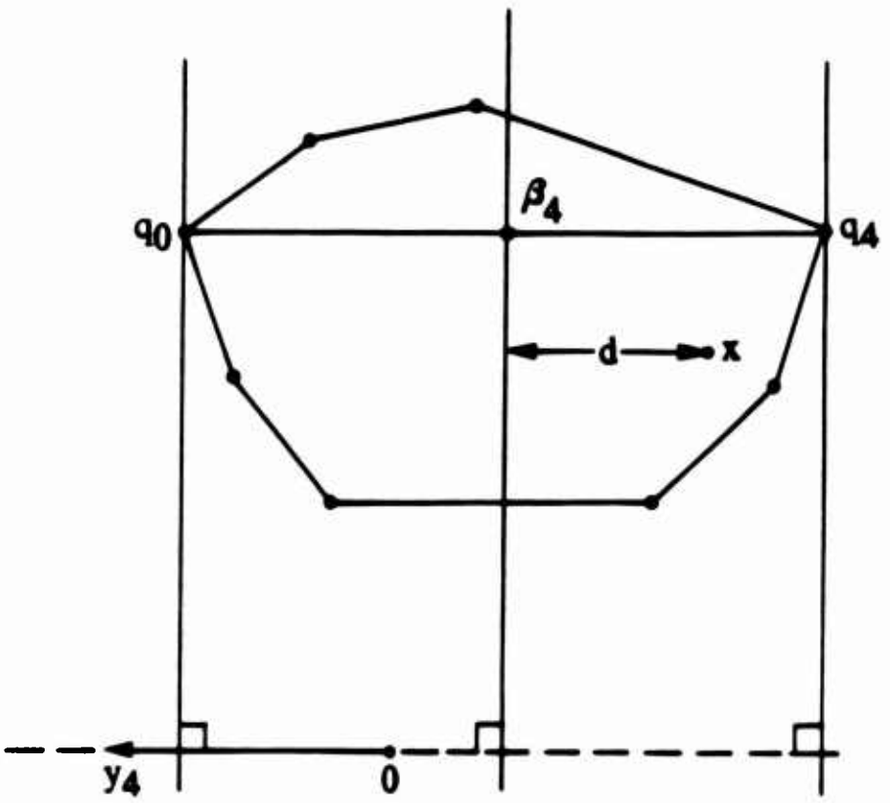



FIGURE 3-7. THE DISTANCE FROM AN ARBITRARY POINT x IN D TO THE LINE THROUGH β AND PERPENDICULAR TO $y_n$ is $d = |h_n(x)|/||y_n||$. HERE S = 8, T = 5, AND n = 4.

### 3.2.6. IMPLEMENTATION

The algorithms presented in the last two sections can be implemented with two computer programs. The first program would implement the algorithm in Section 3.2.5. Its input would be a mask and its output would be a reconstruction algorithm. The second program would implement the program in Section 3.2.4. Its input would consist of a reconstruction algorithm and an autocorrelation function and its output would be the object function. With this arrangement, if one wished to reconstruct many object functions using the same mask, the first program would have to be run only once.

### 3.2.7. CONCLUSIONS

It has been shown that if a function is zero outside a given mask and is non-zero at the reference points of the mask, then it is uniquely determined (up to multiplication by a complex number with modulus 1) by its autocorrelation function among all other object functions which are zero outside the mask. (A mask is a set of points in the discrete lattice whose convex hull has no parallel sides.) Moreover, there is an algorithm for generating reconstruction algorithms for any given mask which in turn can be used to reconstruct object functions satisfying the above mentioned conditions from their autocorrelation functions.

This theory has some similarity to holography [3.10, 3.11]. However, here several (at least 3) reference points are used whereas

only one reference point is needed in the holographic situation. On the other hand, the holographic reference point must be isolated from the rest of the object whereas no such isolation of the reference points is required here. It is interesting to speculate whether there might be a more general theory of which this theory and holography would both be special cases.

## 3.3 EXPERIMENTAL CLOSED-FORM RECONSTRUCTION RESULTS

Autocorrelation data was computer-simulated, including the effects of noise, and images were reconstructed using the closed-form reconstruction algorithm described in the previous section.

For each reconstruction experiment an object, $f(x,y)$, fitting within a triangular support was Fourier transformed:

$$F(u,v) = \mathcal{F}[f(x,y)]$$

The squared modulus, $|F(u,v)|^2$, of the Fourier transform was computed, and it was scaled in intensity so that the total integrated intensity became equal to a given number of photons,

$$N_p = \sum_{uv} |F(u,v)|^2 .$$

Then each intensity sample $|F(u,v)|^2$ was replaced with a random number, $|F_n(u,v)|^2$ drawn from a Poisson distribution with mean and variance equal to $|F(u,v)|^2$. When $|F(u,v)|^2$ is a large number ($\geq 32$), then a Gaussian approximation to the Poisson distribution is used. This Poisson noise process simulates the effect of photon (shot) noise on the measured Fourier intensity data. The normalized RMS error (NRMSE) of the data is given by

$$\text{Data NRMSE} = \left[ \frac{\sum_{uv} \left( |F_n(u,v)| - |F(u,v)| \right)^2}{\sum_{uv} |F(u,v)|^2} \right]^{1/2}$$

A noisy autocorrelation was computed:

$$r_n(x,y) = \mathcal{F}^{-1}\left[ |F_n(u,v)|^2 \right];$$

and an image, $g_n(x,y)$, was reconstructed using the closed-form reconstruction algorithm. The NRMSE of the reconstructed image is given by

$$\text{Image NRMSE} = \left[ \frac{\sum_{xy} \left| a g_n(x,y) - f(x,y) \right|^2}{\sum_{xy} \left| f(x,y) \right|^2} \right]^{1/2}$$

where a is a constant chosen to minimize the error metric, which accounts for the unknown phase constant associated with $f(x,y)$. It can be shown that the value of a that optimizes the Image NRMSE is

$$a = \frac{\sum_{xy} f(x,y) g^*(x,y)}{\sum_{xy} \left| g(x,y) \right|^2}$$

Examples of images of objects reconstructed from noisy data, for which the object is a uniform triangle, are shown in Figures 3-8 to 3-10 for various sizes of the object. Figure 3-11 plots the image NRMSE versus the data NRMSE for the images shown in Figures 3-8 and 3-9. Several interesting effects are evidenced from these reconstruction examples. First, the closed-form algorithm is very sensitive to noise. A fraction of a percent error in the data results in several percent error in the image. Second, increased data error results in increased image error, but only in an average sense. Occasionally the image error can be greater when the data error is less, because for a given amount of data error the image error that one gets is highly variable. Depending on the particular realization of the noise in the data, the three corner points will have different amounts of error. Small differences in the error of the corner points can yield large differences in the error of the image since the corner points are used over and over again and the error from them propagates and is magnified as the recursive steps build upon one another. This also gives rise to a third effect: the error for the interior points of the reconstructed image is much worse than the

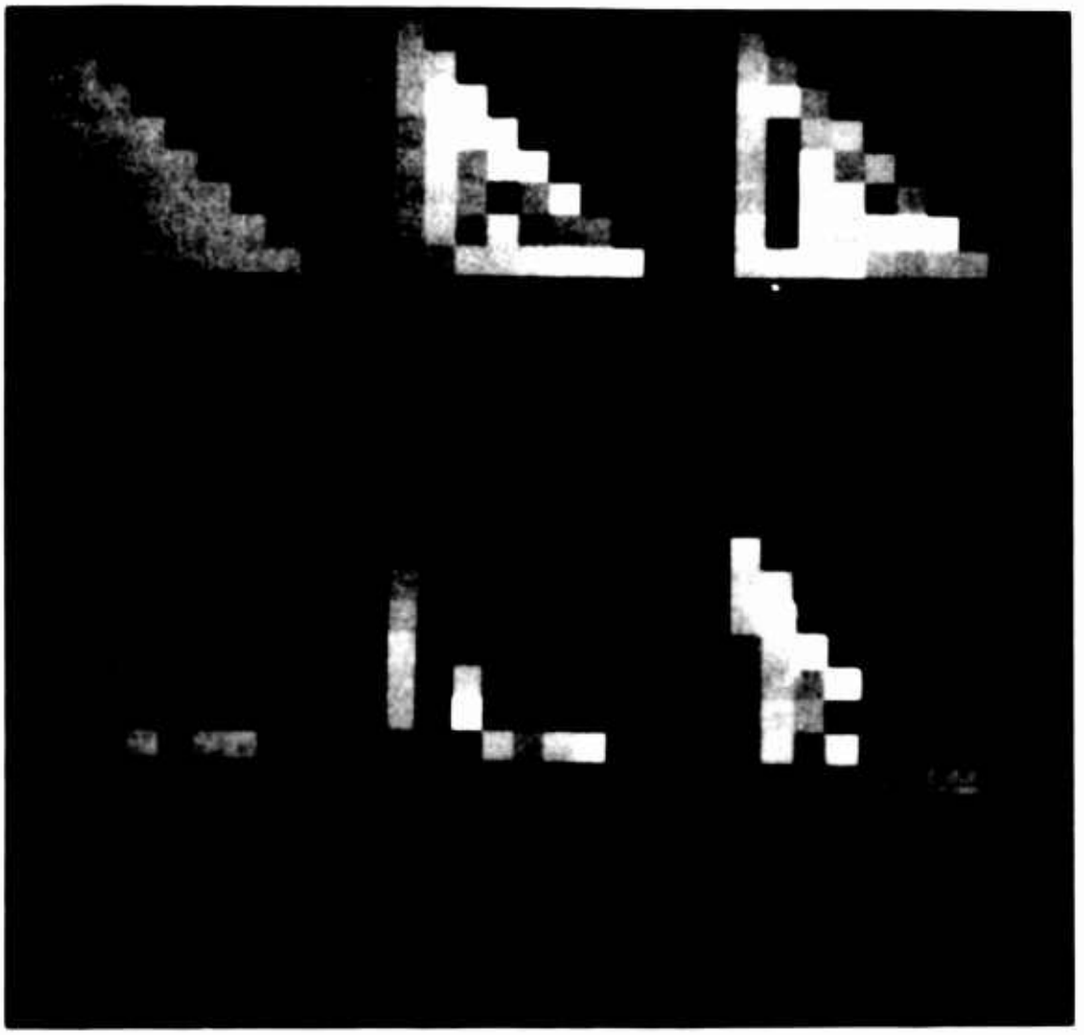


FIGURE 3-8. 8 × 8 OBJECT AND IMAGES RECONSTRUCTED BY THE CLOSED-FORM RECURSIVE ALGORITHM. Number of photons listed are in the Fourier intensity domain.

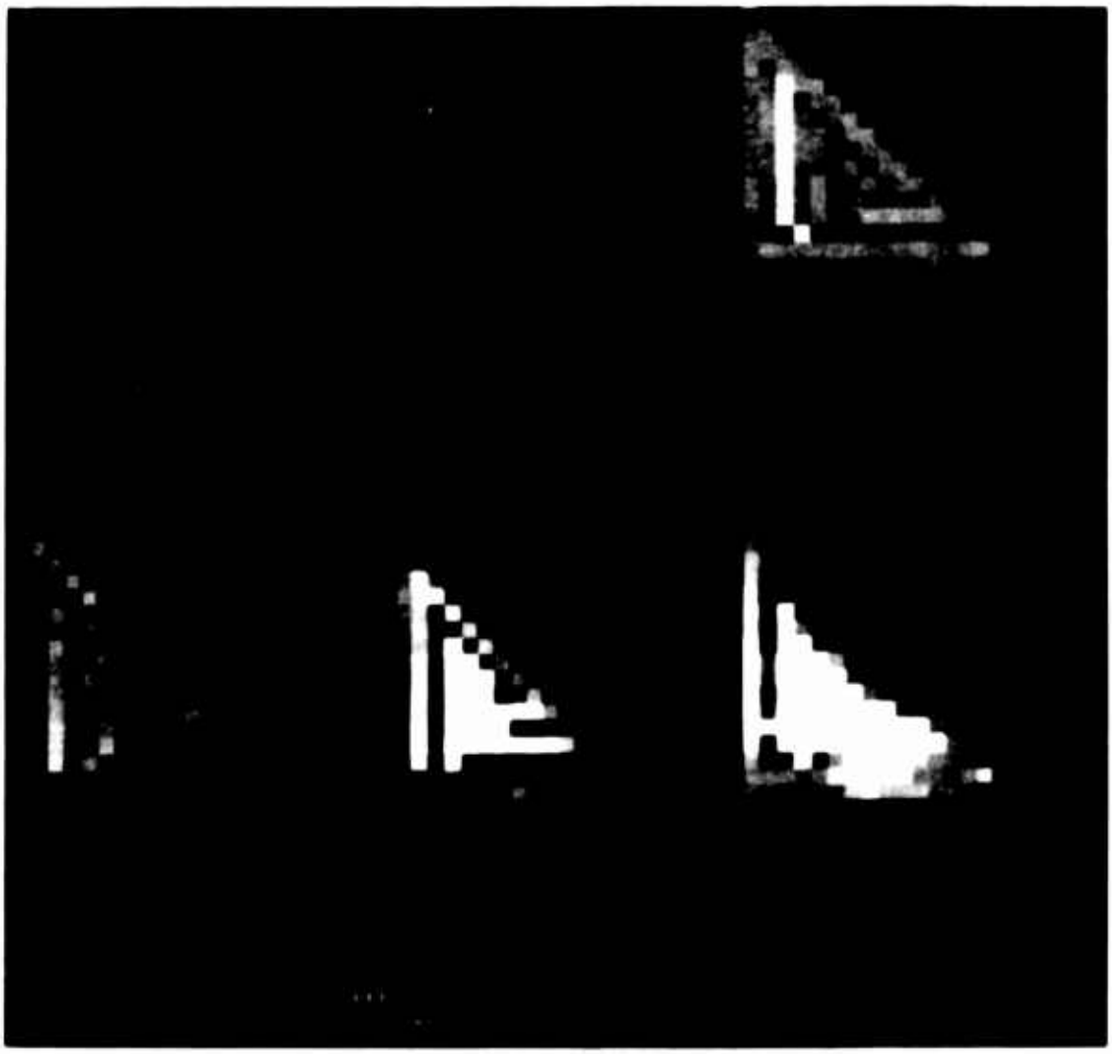


FIGURE 3-9. 16 x 16 TRIANGULAR OBJECT AND IMAGES RECONSTRUCTED BY THE CLOSED-FORM RECURSIVE ALGORITHM

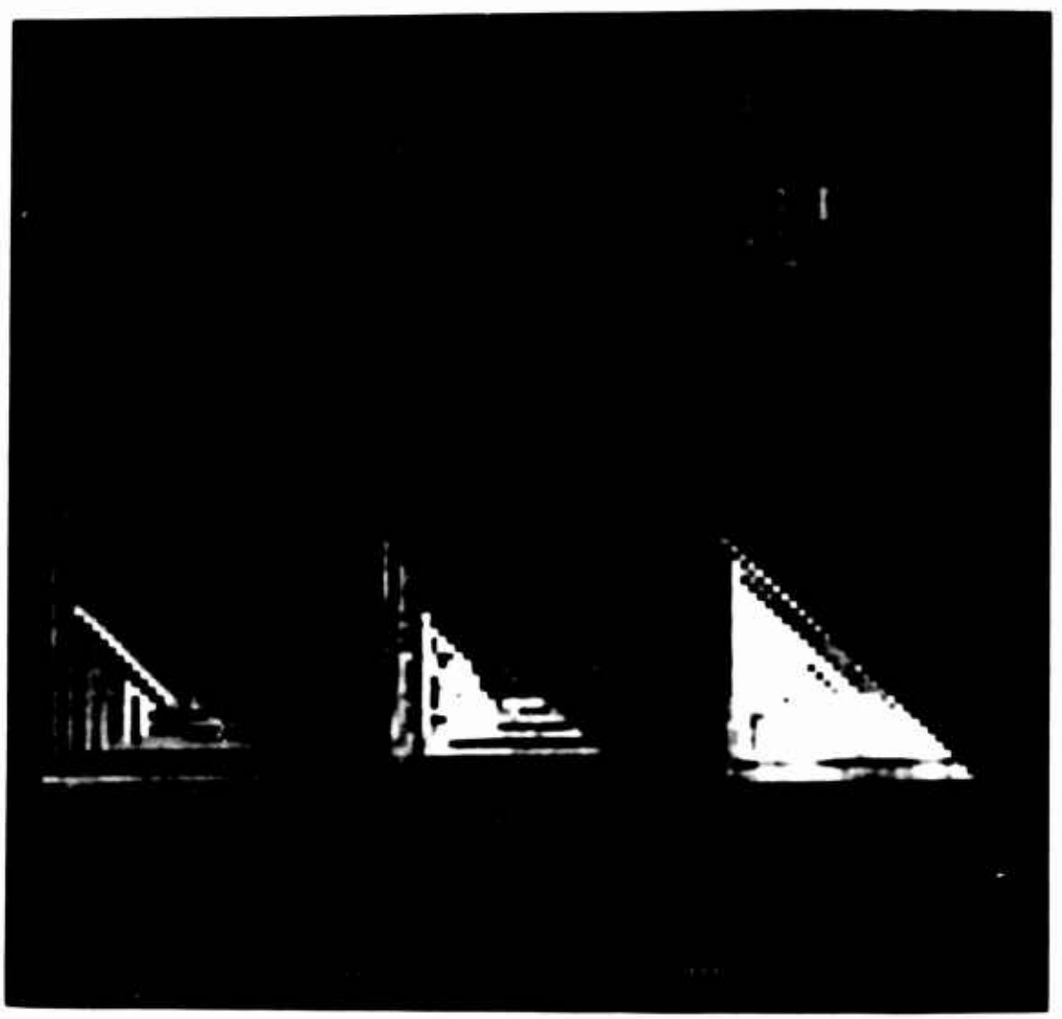


FIGURE 3-10. 32 x 32 TRIANGULAR OBJECT AND IMAGES RECONSTRUCTED BY THE CLOSED-FORM RECURSIVE ALGORITHM

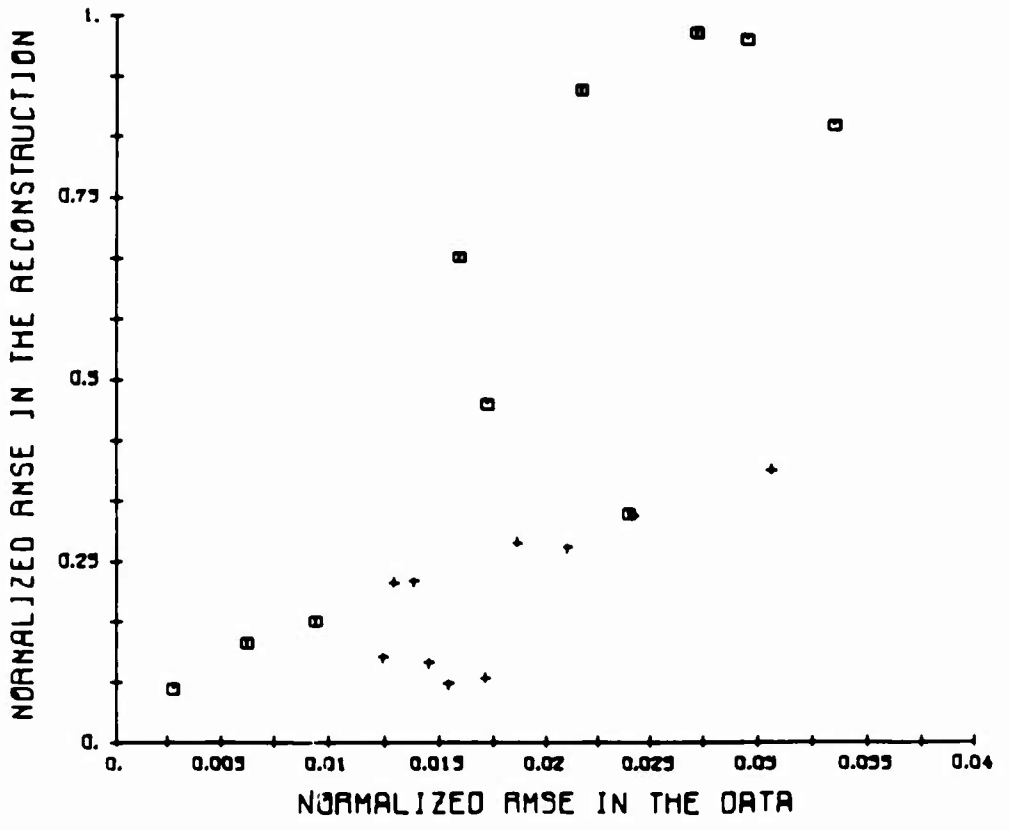



FIGURE 3-11. NRMS ERROR OF THE RECONSTRUCTED IMAGE VERSUS NRMS ERROR OF THE DATA. Crosses are for the 8 x 8 triangles and squares for the 16 x 16 triangles.

error of the edge points in the image. Fourth, for similar reasons the error of larger images is far worse than the error of smaller images. Fifth, a stripe pattern parallel to each of the edges tends to occur. This happens for the following reason. Suppose that one of the three reconstructed corner points is brighter than it should be. Then the opposing edge of the image, the computation of which involves division by the corner point, will tend to be too dark. Then the next inward row (or column) from the edge, the computation of which involves subtraction of terms involving the edge, will tend to be too bright, etc.

How badly this striping effect affects the interpretability of an image was tested by using a picture of an airplane as the object, imbedded in a 32 by 32 triangle. The results of reconstruction experiments from noisy data using the closed-form reconstruction algorithm for the object are shown in Figures 3-12, 3-13 and 3-14. As seen from Figure 3-12, the image can still be discerned through the partially-obscuring striped pattern. Therefore the intelligibility of the image may be understated by the image NRMSE. Figure 3-13 shows the image NRMSE versus the total number of photons for all the noise values tried for this object, and Figure 3-14 shows the same information, but as a function of data NRMSE.

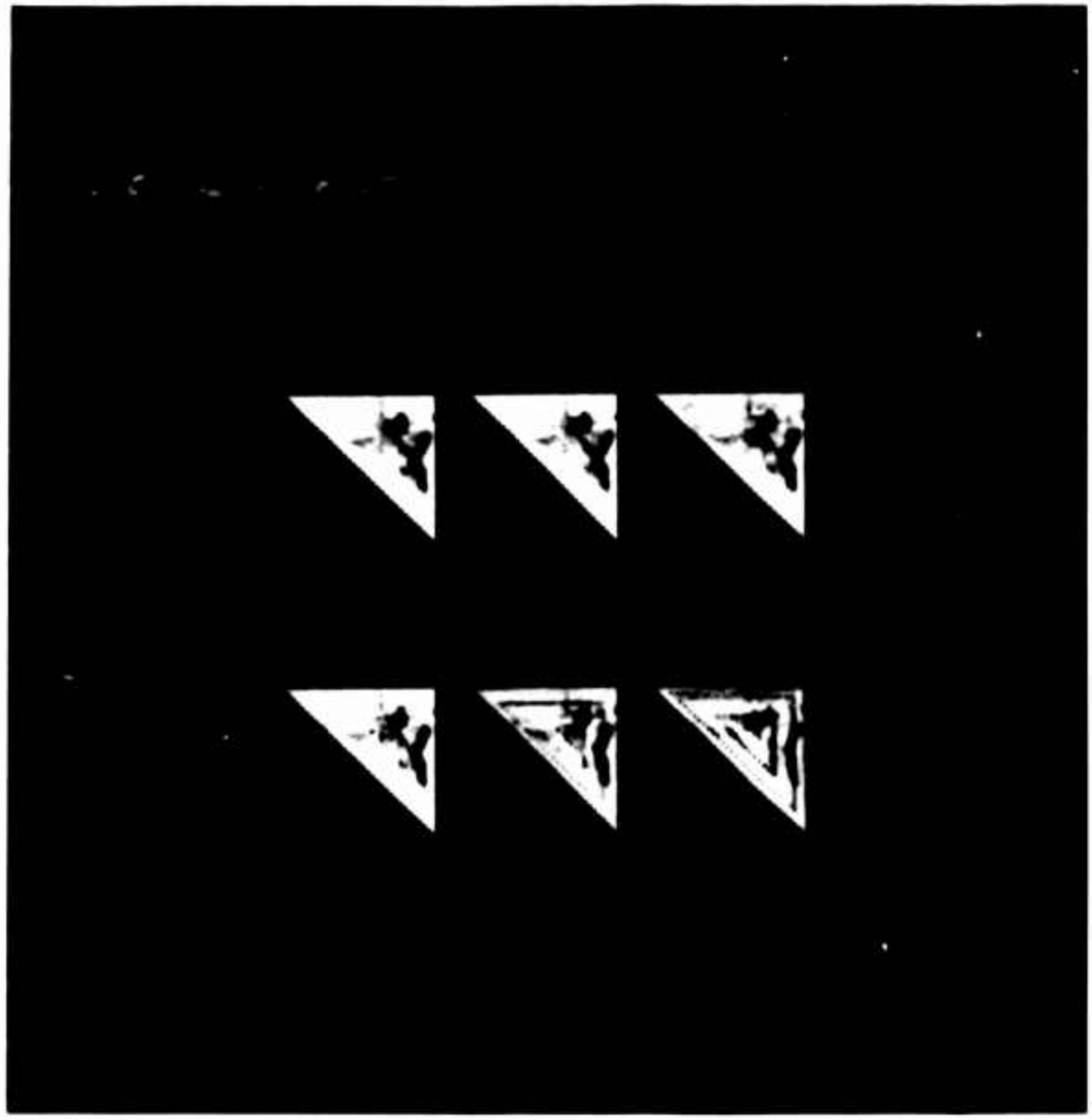


FIGURE 3-12. 32 x 32 JET OBJECT AND IMAGES RECONSTRUCTED BY THE CLOSED-FORM RECURSIVE ALGORITHM. The number of photons in the Fourier intensity data, the corresponding data NRMS error and the reconstructed image NRMS errors are as follows:

| | No. of Photons | Data NRMSE | Image NRMSE |
|---|---|---|---|
| (a) | - Original Object - | | |
| (b) | $10^9$ | 0.0011 | 0.03 |
| (c) | $10^8$ | 0.0034 | 0.10 |
| (d) | $6 \times 10^7$ | 0.0043 | 0.11 |
| (e) | $2 \times 10^7$ | 0.0073 | 0.27 |
| (f) | $10^7$ | 0.0107 | 0.60 |

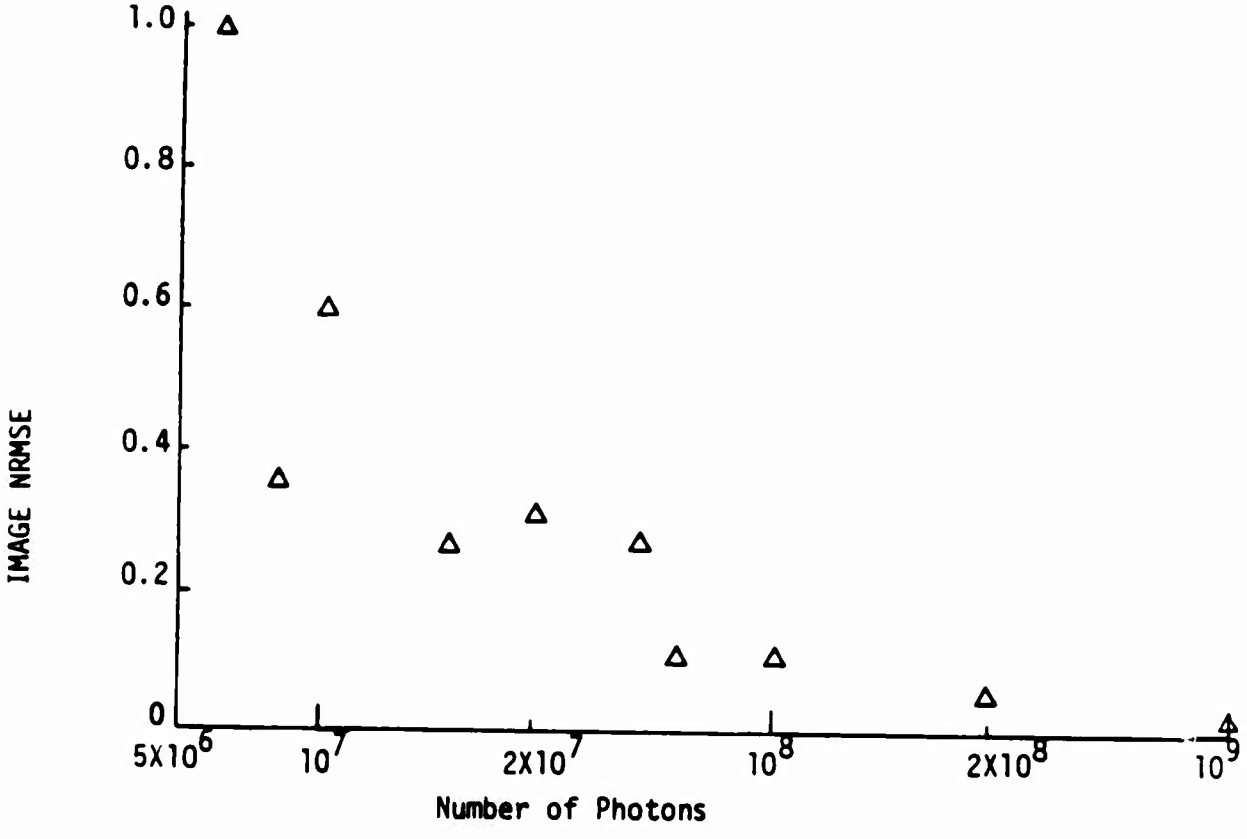



FIGURE 3-13. RECONSTRUCTED JET IMAGE NRMS ERROR VERSUS NUMBER OF PHOTONS

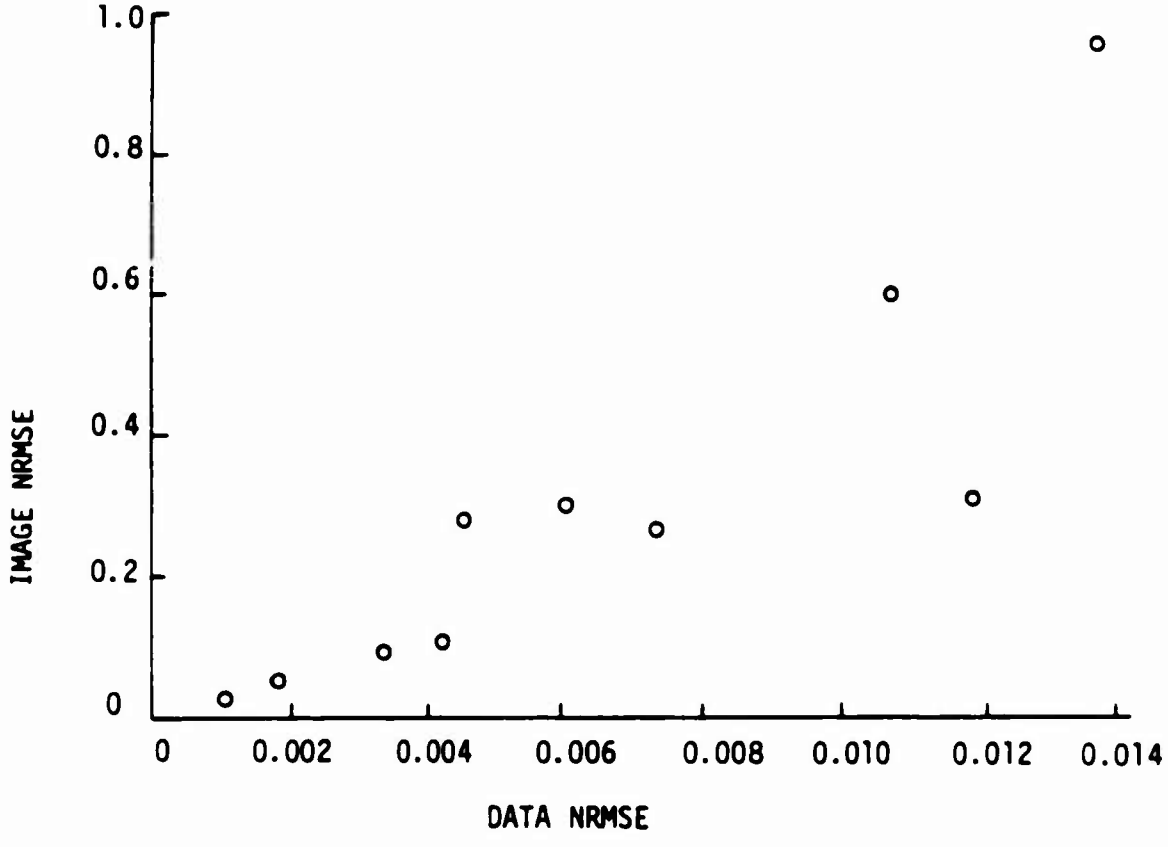



FIGURE 3-14. RECONSTRUCTED JET IMAGE NRMS ERROR VERSUS DATA NRMS ERROR

## 3.4 QUASI-SAMPLING ILLUMINATION PATTERN

One problem with the closed-form reconstruction algorithm described in Section 3.2 is its reliance on a sampled object. In this section, it is shown that by a special kind of illumination one can approximately achieve the desired sampling effect in the target area.

If one illuminates a target area with four mutually coherent point sources in the far field at a distance R given by

$$\delta(u - u_0, v - v_0) + \delta(u - u_0, v + v_0) + \delta(u + u_0, v - v_0)$$

$$+ \delta(u + u_0\ v + v_0)$$

one gets a field at the target area given by the sum of four plane waves (Fourier transforms of the delta functions):

$$(\lambda R)^{-1}\left\{\exp\left[\frac{-i2\pi}{\lambda R}(u_0x + v_0y)\right] + \exp\left[\frac{-i2\pi}{\lambda R}(u_0x - v_0y)\right]\right.$$

$$\left. + \exp\left[\frac{-i2\pi}{\lambda R}(-u_0x + v_0y)\right] + \exp\left[\frac{-i2\pi}{\lambda R}(-u_0x - v_0y)\right]\right\}$$

$$= (\lambda R)^{-1}\left[\exp\left(\frac{i2\pi}{\lambda R}u_0x\right) + \exp\left(\frac{-i2\pi}{\lambda R}u_0x\right)\right]\cdot\left[\exp\left(\frac{i2\pi}{\lambda R}v_0y\right) + \exp\left(\frac{-i2\pi}{\lambda R}v_0y\right)\right]$$

$$= 4(\lambda R)^{-1}\cos\left(\frac{2\pi u_0x}{\lambda R}\right)\cos\left(\frac{2\pi v_0y}{\lambda R}\right)$$

which has intensity

$$16(\lambda R)^{-2}\cos^2\left(\frac{2\pi u_0x}{\lambda R}\right)\cos^2\left(\frac{2\pi v_0y}{\lambda R}\right)$$

which has lines of zeros along $x = \lambda R(n + 1/2)/(2u_0)$ and along $y = \lambda R(n + 1/2)/(2u_0)$, for $n = 0, \pm1, \pm2, \ldots$. This is illustrated in Figure 3-15.

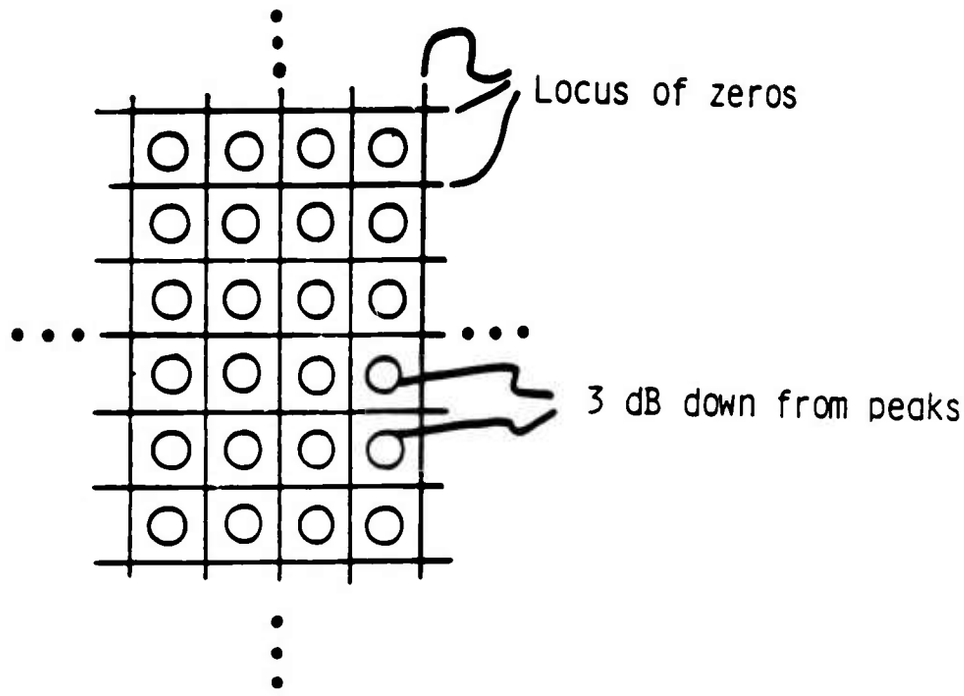



FIGURE 3-15. QUASI-SAMPLING ILLUMINATION PATTERN. The circles are the locus of points 3 dB down from the array of peaks and the lines are the locus of zeros.

The illumination pattern would be of limited extent which could be modeled by multiplying the above by a slowly varying weighting function defining the field-of-view, $(\lambda R/4)t(x,y)$, so that the entire illumination pattern is

$$w(x,y) = t(x,y) \cos\left(\frac{2\pi u_o x}{\lambda R}\right) \cos\left(\frac{2\pi v_o y}{\lambda R}\right),$$

where there are a number of cycles of the cosines over the extent of $t(x,y)$.

What is accomplished by this is a quasi-sampling of the object. It may be possible to use the closed-form recursive reconstruction algorithm to reconstruct an object, illuminated by $w(x,y)$ above, from its autocorrelation, or it may be necessary to make modifications to reduce the errors due to approximating this pattern by a true sampling pattern.

Note that by the addition of more plane waves it is possible to get sharper local maxima and broader stripes of low intensity, but at the expense of a more complicated illumination system with phase-stability problems of its own.

In the real world, the four mutually coherent illumination sources may have an unknown relative phasing between them. If the constant relative phases of the four sources are $\phi_1$, $\phi_2$, $\phi_3$ and $\phi_4$, then the product of cosines like the equation above occurs only if $\phi_1 - \phi_2 - \phi_3 + \phi_4 = 0$. This implies a stringent stability requirement on the illumination system, but it requires the control of only a single parameter (one piston term) rather than the control of the phase of an entire large aperture.

References


3.1. J.R. Fienup, T.R. Crimmins and W. Holsztynski, "Reconstruction of the support of an object from the support of its autocorrelation function," J. Opt. Soc. Am. 72, 610–624 (1982).

3.2. M.H. Hayes and T.F. Quatieri, "The importance of boundary conditions in the phase retrieval problem," IEEE Trans. Acoust. Speech Signal Process. ASSP-82, 1545 (1982).

3.3. Yu. M. Bruck and L.G. Sodin, "On the ambiguity of the image reconstruction problem," Opt. Commun. 30, 304-308 (1979).

3.4. M.A. Fiddy, B.J. Brames and J.C. Dainty, "Enforcing irreducibility for phase retrieval in two dimensions," Opt. Lett. 8, 96-98, (1983).

3.5. J.R. Fienup, "Phase retrieval algorithms: a comparison," Appl. Opt. 21, 2758-2769 (1982).

3.6. J.R. Fienup, "Reconstruction of an object from the modulus of its Fourier transform," Opt. Lett. 3, 27-29 (1978).

3.7. J.R. Fienup, "Space object imaging through the turbulent atmosphere," Opt. Eng. 18, 529-534 (1979).

3.8. J.R. Fienup, "Reconstruction of objects having latent reference points," J. Opt. Soc. Am. 73, 1421-1426 (1983).

3.9. B.J. Brames, "Unique phase retrieval with explicit support information," Opt. Lett. 11, 61-63 (1986).

3.10. J.W. Goodman, "Analogy between holography and interferometric image formation," J. Opt. Soc. Am. 60, 506-509 (1970).

3.11. C.Y.C. Liu and A.W. Lohman, "High resolution image formation through the turbulent atmosphere," Opt. Commun. 8, 372-377 (1973).

# 4
# CONSTRAINT INVESTIGATION

In this section the various forms of constraints that might be useful for phase retrieval are discussed. Section 4.1 describes results obtained with a variety of support (illumination pattern) constraints. The vast majority of the effort to date concentrated on developing imaging concepts based on the support constraint. Section 4.2 describes other constraints that might also prove useful.

## 4.1 EFFECT OF ILLUMINATION PATTERN SHAPE

The support, S, of an object is defined as the set of points for which the object is nonzero. For the case of a satellite imaged against the night sky or a ship imaged on calm water with a SAR, the support of the object is basically the filled-in outline of the object. For an airborne or spaceborne sensor looking downward at a general scene, the extent of the object is basically defined by the field-of-view of the sensor. This latter case does not represent a useful support constraint. However, for an imaging system employing active illumination, the transmitted beam (the illumination beam) can take on a known shape at the plane of the target, and it can be designed to occupy an area smaller than the field-of-view of the receiver. Then the effective support of the object is the support of the illumination beam pattern. For the case of a SAR, it is assumed that when no phase is available the pulse repetition frequency is at least twice that ordinarily required by Nyquist sampling when phase information is available.

The two most important properties of an illumination pattern are its shape (elliptical, rectangular, polygonal, etc.) and its taper (how slowly it transitions from the bright part of the pattern to where it is effectively zero). As shown in the proposal [4.1, p. 2-29], phase retrieval algorithms are much more effective for some shapes (which we refer to as strong shapes) than for others. Furthermore, phase

retrieval algorithms are more effective for sharp support constraints, i.e. when there is little or no taper to the illumination pattern [4.1, p. 2-25]. Section 5 of this report details the results of our investigation of the effects of tapered illumination patterns and algorithm improvements that were made for the case of larger amounts of taper. In what immediately follows we discuss the effects of the shape of the support.

In early phase retrieval work there was not an awareness that the support of the object played an important role in the success of phase retrieval. Early successful reconstruction results were for space objects whose supports were naturally non-centrosymmetric [4.2]. Other groups attempted phase retrieval for unnatural objects -- scenes bounded by squares -- and were unsuccessful. Fiddy, Brames and Dainty [4.3] found that the iterative Fourier transform algorithm, although it worked poorly for a rectangular support, worked well for a support consisting of a rectangle plus an extra point just off one corner of the rectangle. This latter support has the special property that any sampled function defined on that support, which is nonzero at the extra point and at one opposite corner, has a Fourier transform that is a nonfactorable polynomial according to Eisenstein's irreducibility theorem. This implies that the phase retrieval problem in unconditionally unique for objects of this type. In retrospect, from those results we can make the crucial connection between three different aspects of the phase retrieval problem: the support of the object, the uniqueness of phase retrieval, and the success of the iterative Fourier transform algorithms.

The trends connecting those three elements, which we have continued to confirm, are the following. First, the support of the object determines whether ambiguities are possible. Second, objects for which uniqueness can be proven are easier to reconstruct by the iterative Fourier transform algorithm than are other objects. The first trend is amply demonstrated in Section 3.2 which shows that sampled objects

having known, convex hulls with no parallel sides are unique. The second trend is shown by the reconstruction results [4.1, pp. 2-24 and 2-28] in which objects having known triangular support (which are unique -- see Section 3.2) and objects having known supports with separated parts (which even in one dimension are usually unique -- see examples in Section 5) are easily reconstructed while objects with other support constraints, like that of a single ellipse or a single rectangle are difficult to reconstruct.

The closed-form reconstruction algorithm described in Section 3.2 may not be practical for use on real-world data, since it requires the objects to be modelled discretely (as a grid of delta functions or sampled points) and it is very sensitive to noise, particularly if the vertex points are dim (see Section 3.3). Nevertheless, it does constitute a uniqueness proof for the types of objects to which it can in theory be applied: objects whose support has a convex hull with no parallel sides. This leads us to consider illumination patterns of this type. Figure 4-1 shows an example of a reconstruction experiment using an illumination-pattern shape suggested by the uniqueness proof. On the left is the modulus of a complex-valued SEASAT SAR image multiplied by a binary pattern (representing the illumination pattern) in the shape of a pentagon. In the center is the modulus of its Fourier transform (the Fourier phase was discarded). The iterative Fourier transform algorithm was used to reconstruct an image, the modulus of which is shown on the right, from the Fourier modulus using the known support pattern. The result shown is after several hundred iterations (it had not completely converged yet), and it strongly resembles the original object, although not perfectly. Given the difficulty in reconstructing complex-valued images with contiguous supports (with the exception of triangular support) [4.1, pp. 2-24 to 2-30], the success of this kind of support constraint would have been difficult to anticipate were it not for the uniqueness proof.

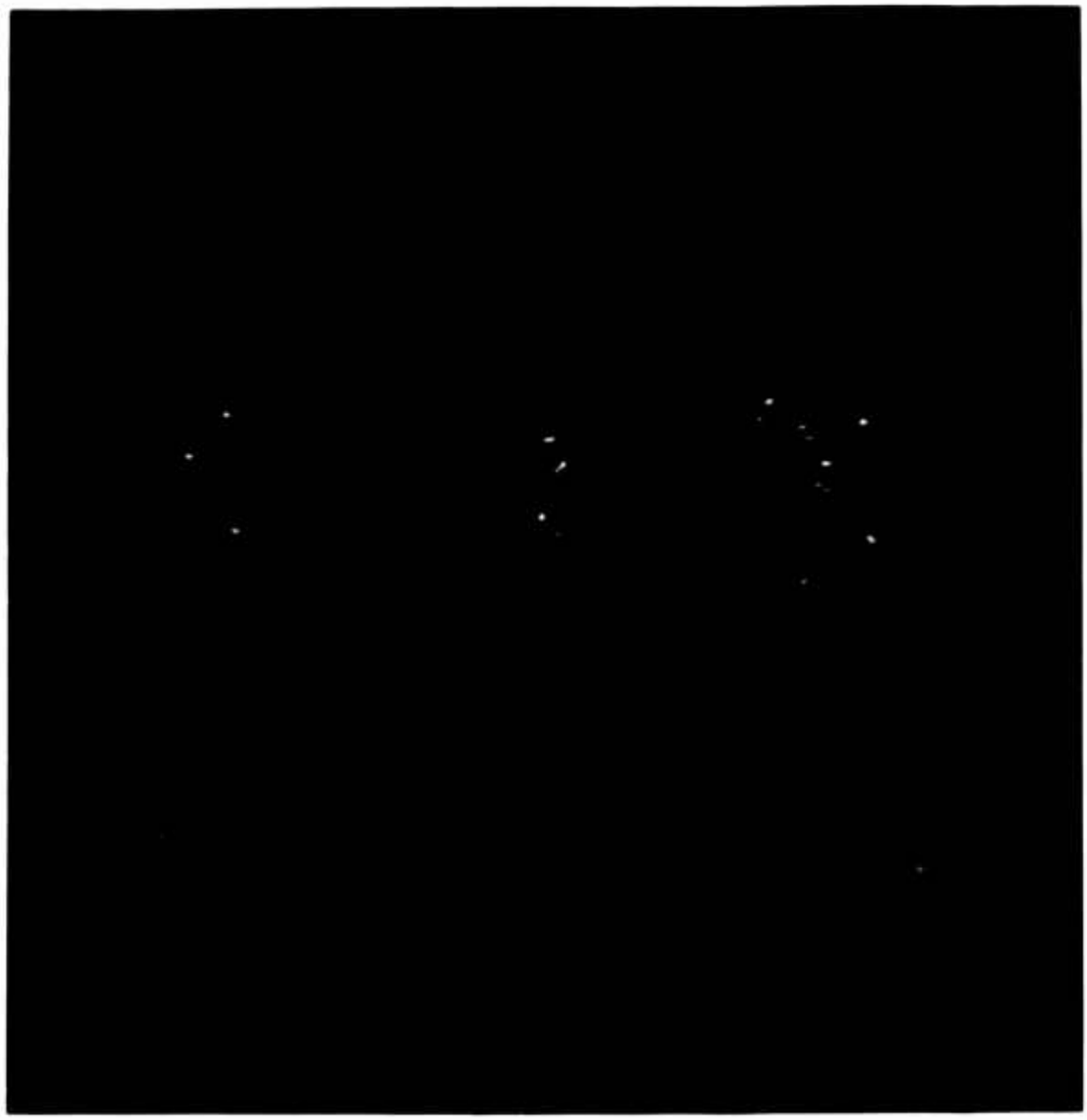


FIGURE 4-1. RECONSTRUCTION EXPERIMENT USING PENTAGON-SHAPED ILLUMINATION PATTERN. Left to right: modulus of complex-valued illuminated SEASAT SAR scene; modulus of signal history (Fourier transform); modulus of image reconstructed from modulus of signal history and pentagon-shaped support constraint using the iterative Fourier transform algorithm.

In summary, to date the support constraints found to be most useful for phase retrieval are (a) supports having two or more well-separated parts and (b) supports having convex hulls with no parallel sides. The farther the sides are from being parallel, the better.

## 4.2 OTHER CONSTRAINTS

Any information that is in a domain other than the domain of the measured data has the potential for being a useful constraint for phase retrieval. Constraints in the domain of the measured data usually just limit the available data and tend not to be useful for phase retrieval. Candidate constraints for the phase retrieval problem are listed in Table 4.1.

Table 4-1
CANDIDATE CONSTRAINTS

Support (illumination pattern)
Nonnegativity
Polarization
Transmitted waveform
Point scatterer in scene
Other scene characteristics

Of these, the support constraint received the most attention, and it is discussed in Section 4.1. The other constraints are described in what follows.

### Nonnegativity

The nonnegativity constraint has been very useful in previous phase retrieval efforts [4.2] and also in other image reconstruction problems, such as tomographic reconstruction from incomplete projections and constrained deconvolution. Unlike the support constraint, nonnegativity

must exist naturally -- we do not know how to impose it artificially. It naturally occurs in most passive, noncoherent imaging scenarios. The brightness distribution of an incoherently-illuminated reflecting object or a self-emissive object is characterized by a real, nonnegative function (power or photons per unit area). This can be true for actively illuminated objects as well, as long as the illumination is sufficiently incoherent. An exception in which this constraint may not be valid is for passive doppler imaging as encountered in the PACE program [4.4], for which the aperture function is band-pass. Nonnegativity is not as useful for band-pass systems since the impulse response has very large negative (or complex-valued) sidelobes which convolve the image, destroying its nonnegativity. For the cases in which nonnegativity naturally does occur, it should be relied on heavily as a phase retrieval constraint.

Polarization

Certain kinds of reflecting objects have distinctly different reflectivities for the two different receive polarities (i.e. for same polarity as transmit or for opposite polarity). As an example, corner reflectors reflect either very strongly or very weakly depending on the polarization. Unfortunately, from a single collection it is not immediately obvious how to utilize this information. On the other hand, if two collections are made simultaneously, one for each polarization, then there is increased possibility of using polarization advantageously. One such possibility would be to use the difference between two degraded images (with measured Fourier phase in the presence of phase errors) to identify point-like reflectors Then the point-like reflectors could be used in the prominent-point processing described below. Other examples of using polarization may be also be possible.

Transmitted Waveform Type

Early on in the program it was thought that perhaps transmitting a pulse with missing frequency bands might be useful insofar as it would constitute a support-like constraint in the signal history. Upon further examination, it appears that such missing frequencies would primarily result in a loss of data rather than constituting a useful constraint.

A point worth making relating to transmitted wavefront is that the use of phase retrieval techniques may facilitate the use of nonconventional waveforms. As the transmitted waveform departs from the standard set (e.g. the chirp waveform), the availability of hardware that can form the desired waveform in a phase-stable manner may be questionable. By reducing the tolerance on the phase stability of a waveform generator it may be possible to achieve waveforms that would otherwise be very difficult to produce. The reduced tolerance imaging/phase retrieval techniques may provide the means for reducing the phase stability of the waveform generator while maintaining the desired resolution.

Point Scatterers in Scene

Presently, point-like scatterers (prominent points) in the target area are used for correcting small amounts of phase errors in SAR signal histories [4.5, 4.6]. Prominent point processing can also be of great utility for the case of severe phase errors or when no phase information at all is measured. One particular scenario for phase correction in the presence of large one-dimensional phase errors has already been demonstrated [4.6]. For the case of motion compensation errors in SAR, one has a one-dimensional (azimuth) phase error. This occurs particularly for the case of inverse SAR, for example the radar is ground-based and the (noncooperative) target flies by with a poorly known flight path and rotation. If there exists a dominant prominent

point scatterer in a given compressed range cell, then it can be used to calculate the azimuth phase error (taking its phase to be the phase error). The phase errors in all range bins can be corrected by subtracting that phase.

Other Scene Characteristics

The constraints mentioned above are common to large classes of imagery. Also, there may often be additional constraints that exist in specific instances. For example, if the scene has been imaged by another sensor system or by a similar sensor at an earlier time, then these additional images may contain information that can be counted on to appear in the present image and therefore can be used as an a priori constraint. Examples include the known existence of permanent cultural targets or of no-return areas such as lakes or smooth surfaces.

## REFERENCES


4.1 "Reduced Tolerance Imaging," ERIM Proposal 653143 to DARPA/TTO, May 1984.

4.2 J.R. Fienup, "Space Object Imaging through the Turbulent Atmosphere," Opt. Eng. 18, 528-534 (1979).

4.3 M.A. Fiddy, B.J. Brames, and J.C. Dainty, "Enforcing Irreducibility for Phase Retrieval in Two Dimensions," Opt. Lett. 8, 96-98 (1983).

4.4 K.K. Ellis, I.J. LaHaie, A.M. Tai, N. Subotic and I. Cindrich, "Passive Interferometric Imaging," Technical Report to AF Wright Aeronautical Labs., Contract No. F33615-84-C-1508 (December 1985).

4.5 J.R. Fienup, "Digital Focusing" in D.E. Klingler et al., "Advanced Synthetic Array Radar Techniques (U)," Third Interim Report, AFAL-TR-77-83, November 1977, pp. 204-233 (SECRET).

4.6 J.C. Dwyer, J.R. Fienup and I. Cindrich, "Hybrid-Optical Prominent Point Processing of Radar Data (U)," 26th Annual Tri-Service Radar Symposium Record, July 1980, p. 285 (SECRET).

# 5
# RECONSTRUCTION OF OBJECTS WITH TAPERED ILLUMINATION

## 5.1 STATEMENT OF PROBLEM

It is well known that knowledge of the support of an object can be a powerful source of information in image-reconstruction problems. By support we mean a compact region outside of which the object is known to be zero, and we denote the set of points that make up the support by the symbol S. In particular, considerable success has been realized in reconstructing an object from its Fourier modulus and a known support [5.1,5.2]. In the reduced-tolerance imaging program an effort is being made to exploit this ability.

Consider an active sensor system that illuminates a target area so that the illumination is confined to a predetermined region. Let $h(x,y)$ be the complex reflectivity of the target:

$$h(x,y) = |h(x,y)|\, e^{i\phi_h(x,y)}. \tag{5-1}$$

Let $w(x,y)$ be the complex illumination function:

$$w(x,y) = |w(x,y)|\, e^{i\phi_w(x,y)}. \tag{5-2}$$

We define the effective object as the product of the complex reflectivity of the target and the illumination function:

$$\begin{aligned} f(x,y) &= w(x,y)\, h(x,y) \\ &= |f(x,y)|\, e^{i\phi(x,y)} \end{aligned} \tag{5-3}$$

The effective object will now have a support corresponding to the known extent of w(x,y). The intensity pattern of the field emanating from the illuminated target is measured in the far field which may be interpreted as the squared modulus of the Fourier transform of the effective object. Known phase-retrieval algorithms may then be employed to reconstruct the effective object from the support constraint and the measured Fourier modulus.

Notice that there is some freedom in the choice of the form of the illumination pattern. For example, the shape of the outline of the pattern could be specifically selected to enhance the usefulness of the support constraint. It is known that certain symmetries in object support can create stagnation problems in phase-retrieval algorithms. Consequently the outline of the illumination pattern should have an asymmetric shape. Furthermore, there is some evidence that a support consisting of disjoint regions can be an advantage in phase retrieval. Finally, it is useful to choose an illumination function with a constant modulus over most of the region of illumination thus facilitating the inversion of Eq. (5-3):

$$h(x,y)\Big|_{(x,y)\in S} = \frac{f(x,y)}{w(x,y)} = \frac{|f(x,y)|}{|w(x,y)|}\, e^{i(\phi(x,y)-\phi_w(x,y))} \tag{5-4}$$

when one desires the complex reflectivity of the target without the influence of the illumination pattern.

Unfortunately, the modulus of the illumination pattern will not be binary in practice, but will have some taper associated with it at the edges, due to the effects of diffraction by the aperture of the illuminator. The contrast between an ideal untapered illumination pattern and a more realistic illumination function is illustrated in

Figure 5-1. Intuitively one might expect, and experimentally it has been shown [5.1,5.2], that the reconstruction of an object from its Fourier modulus and support would be more challenging for an object with a tapered profile than for one with a sharp profile. It was the purpose of this inquiry to explore this issue via computer simulation and to look for algorithmic modifications that would enhance restoration for this case. For example, it was hoped at the outset that any difficulties incurred by tapered illumination might be offset by the support being disjoint.

## 5.2 PRELIMINARY SIMULATIONS

We began by exploring the effect of tapered illumination on phase retrieval by means of computer simulation. A pair of disjoint ellipses was used as the basic shape for the illumination pattern. The untapered illumination pattern was assigned a value of unity within the ellipses and zero outside. Taper was introduced by convolving the binary ellipses with a convolution kernel. The normalized kernels used in these preliminary simulations are shown in Figure 5-2. Cross sections of the edge of the resulting illumination patterns are given in Figure 5-3.

As mentioned earlier, it was speculated that disconnected support might help to overcome any problems associated with tapered illumination. For this reason the total illumination pattern was chosen to be two disjoint ellipses. Simulations were performed for objects with differing amounts of illumination taper and differing amounts of separation between ellipses in the illumination pattern. A given simulation was performed by first multiplying complex SEASAT SAR imagery by the given illumination pattern to create an effective object. Because it is the effective object that we try to recover through phase-retrieval techniques, we will henceforth refer to this as the true object. This object was Fourier transformed with an FFT and the Fourier magnitude was retained. The known region of support in the object

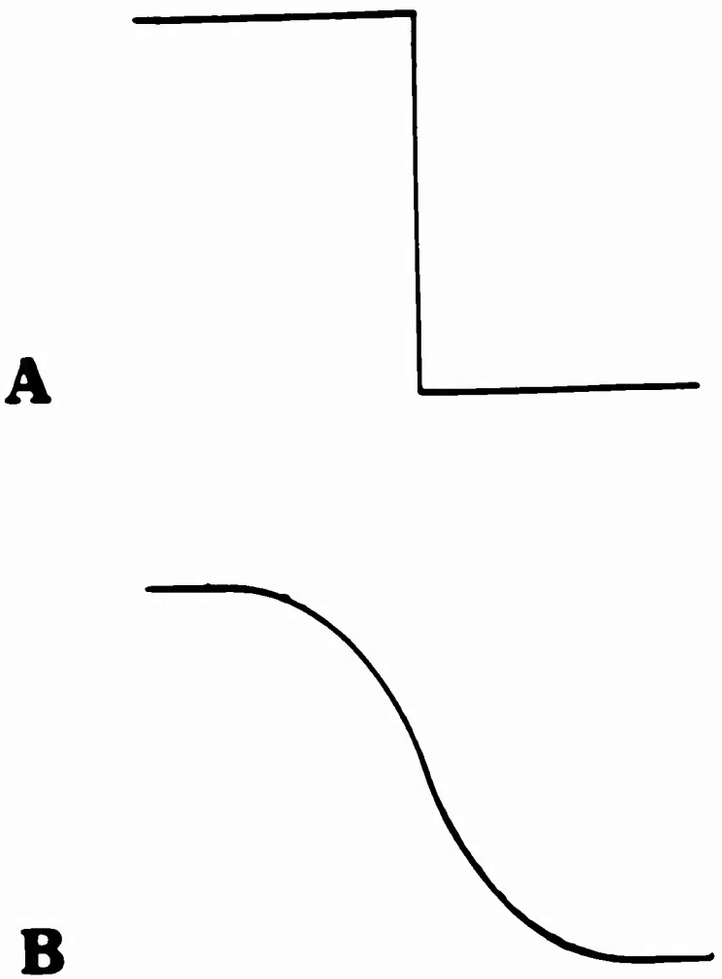



FIGURE 5-1. CROSS SECTIONS OF EDGES OF ILLUMINATION PATTERNS.
A. Ideal binary illumination. B. Tapered illumination.

**A**

| | | |
|---|---|---|
| 1/16 | 1/16 | 1/16 |
| 1/16 | 1/2 | 1/16 |
| 1/16 | 1/16 | 1/16 |

**B**

| | | |
|---|---|---|
| 1/9 | 1/9 | 1/9 |
| 1/9 | 1/9 | 1/9 |
| 1/9 | 1/9 | 1/9 |

FIGURE 5-2. DISCRETE CONVOLUTION KERNELS USED TO ADD TAPER TO BINARY ILLUMINATION PATTERN. A. Center-weighted kernel yields taper #1. B. Evenly-weighted kernel yields taper #2.

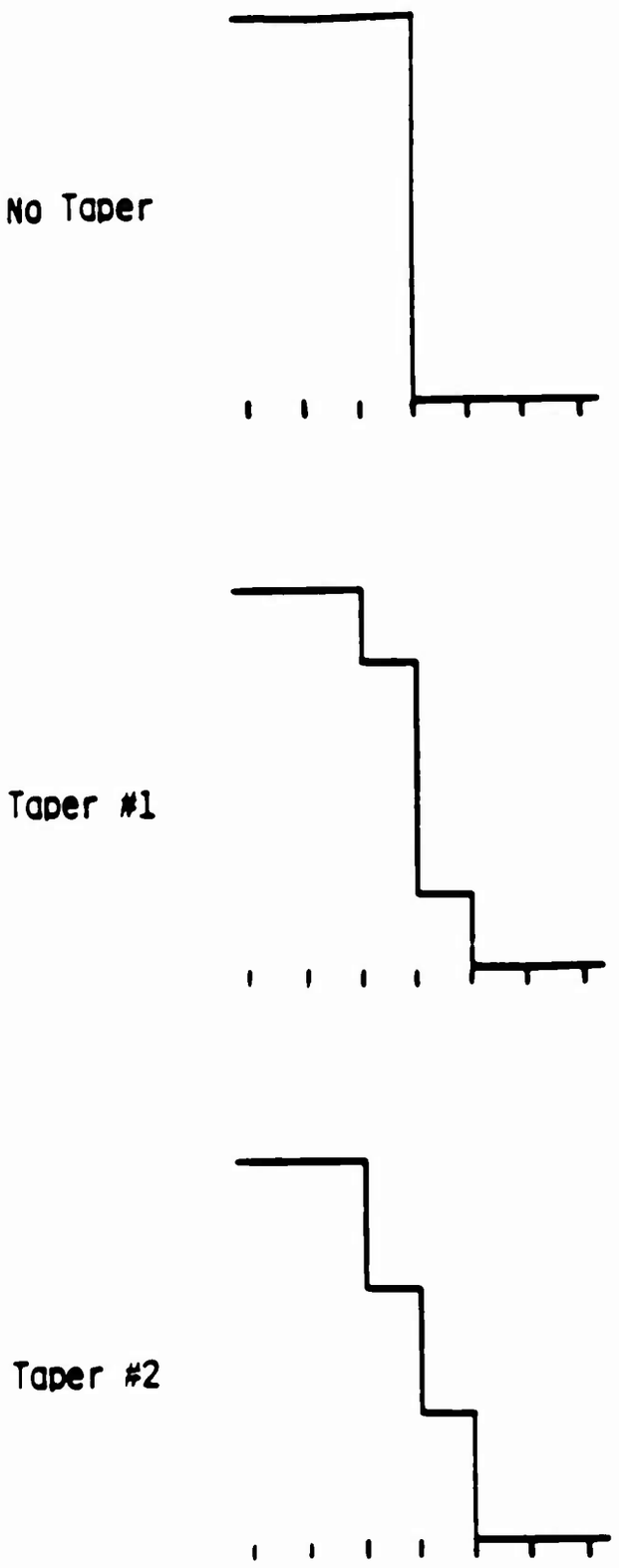



FIGURE 5-3. CROSS SECTIONS OF ILLUMINATION OF TAPER USED IN PRELIMINARY SIMULATIONS

domain was supplied by hard limiting (thresholding) the illumination pattern with a very small threshold value. A standardized sequence of error-reduction and hybrid input-output iterations [5.3] were then performed to reconstruct the object from its Fourier modulus and support. Convergence for all simulations was monitored by calculating a Fourier domain normalized error metric,

$$E_F^2 = \frac{\sum_{u,v} (|G(u,v)| - |F(u,v)|)^2}{\sum_{u,v} |F(u,v)|^2} \qquad (5\text{-}5)$$

where F(u,v) is the discrete Fourier transform of the true object f(x,y) and G(u,v) is the Fourier transform of the image estimate. The convergence is portrayed in Figure 5-4 for six kinds of illumination--three amounts of taper, each with two amounts of separation between ellipses. It is important to note that Figure 5-4 is a log-log plot and therefore the behavior of the algorithm becomes horizontally compressed with increasing number of iterations. Figures 5-5, 5-6, and 5-7 give the final reconstructions for each of the cases tested. These results confirm our expectation that increased amounts of illumination taper make the reconstruction process more difficult. In fact, for the case with the largest amount of taper the algorithm convergence appears to have stagnated. This is in spite of the fact that the amount of taper is extremely mild. There are 51 pixels along the major axis of the large ellipse and only two pixels of taper at the edge. Thus convergence appears to be relatively sensitive to illumination taper. It is important to realize that the convergence curves shown in Figure 5-4 correspond to a specific initial estimate and that the convergence behavior could vary when alternative initial estimates are used.

## 5.3 THE SHRUNKEN-MASK ALGORITHM

In order to explore the reasons for stagnation we created a difference image between the modulus of the true object and that of the restored object for the case of intermediate taper (taper #1). This

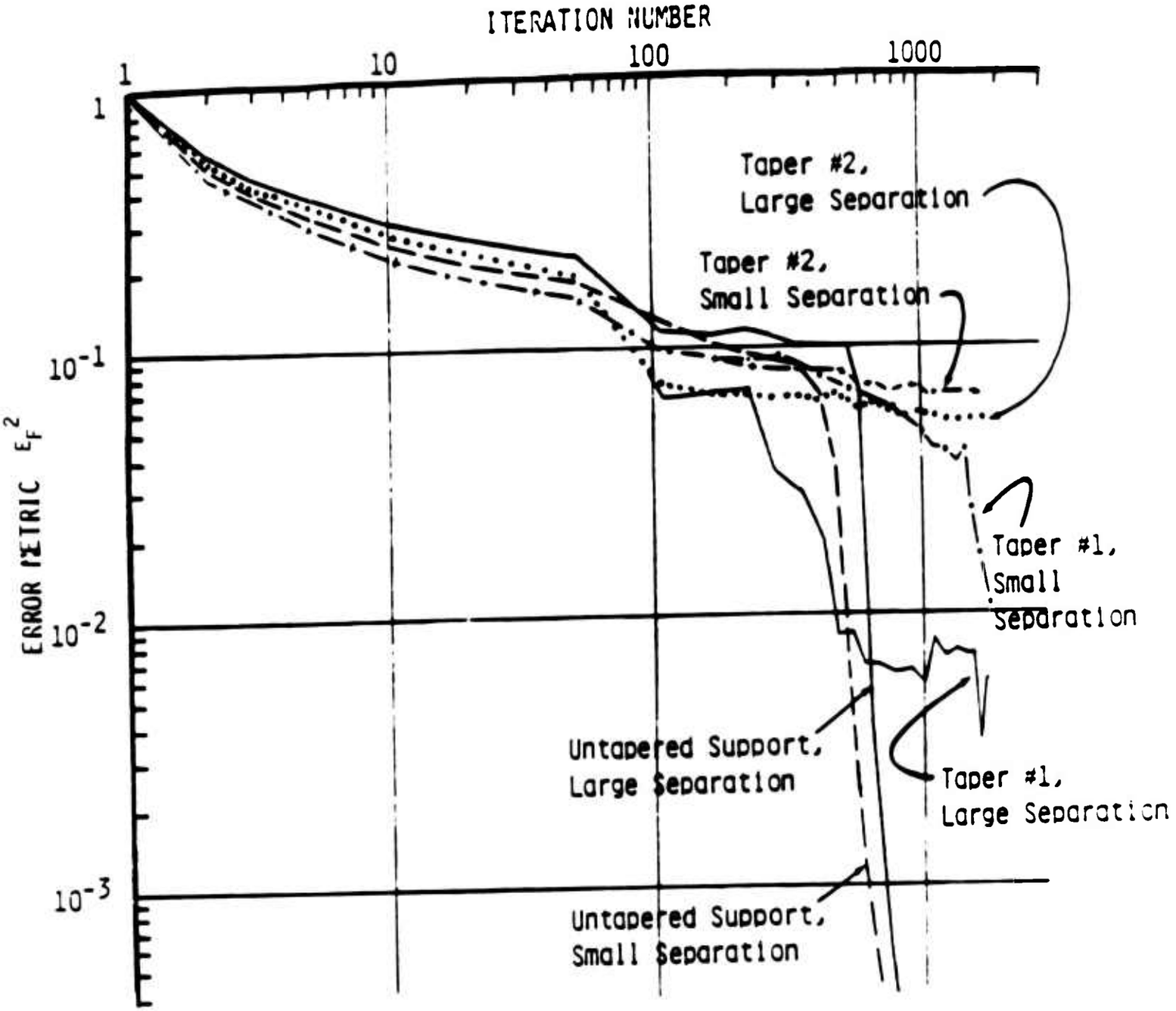



FIGURE 5-4. CONVERGENCE BEHAVIOR AS A FUNCTION OF ILLUMINATION TAPER AND SUPPORT SEPARATION

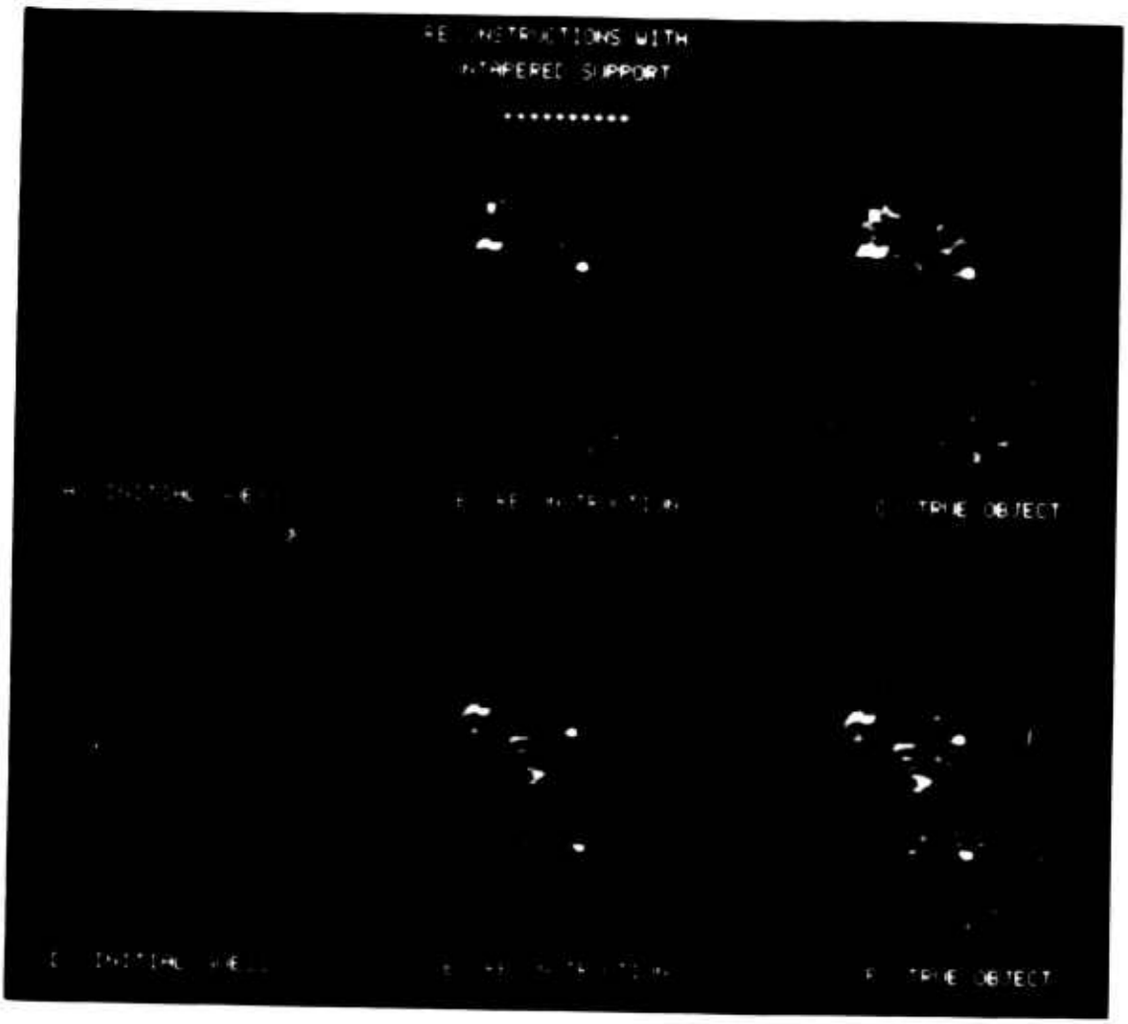



FIGURE 5-5. RECONSTRUCTIONS OF OBJECTS WITH UNTAPERED ILLUMINATION

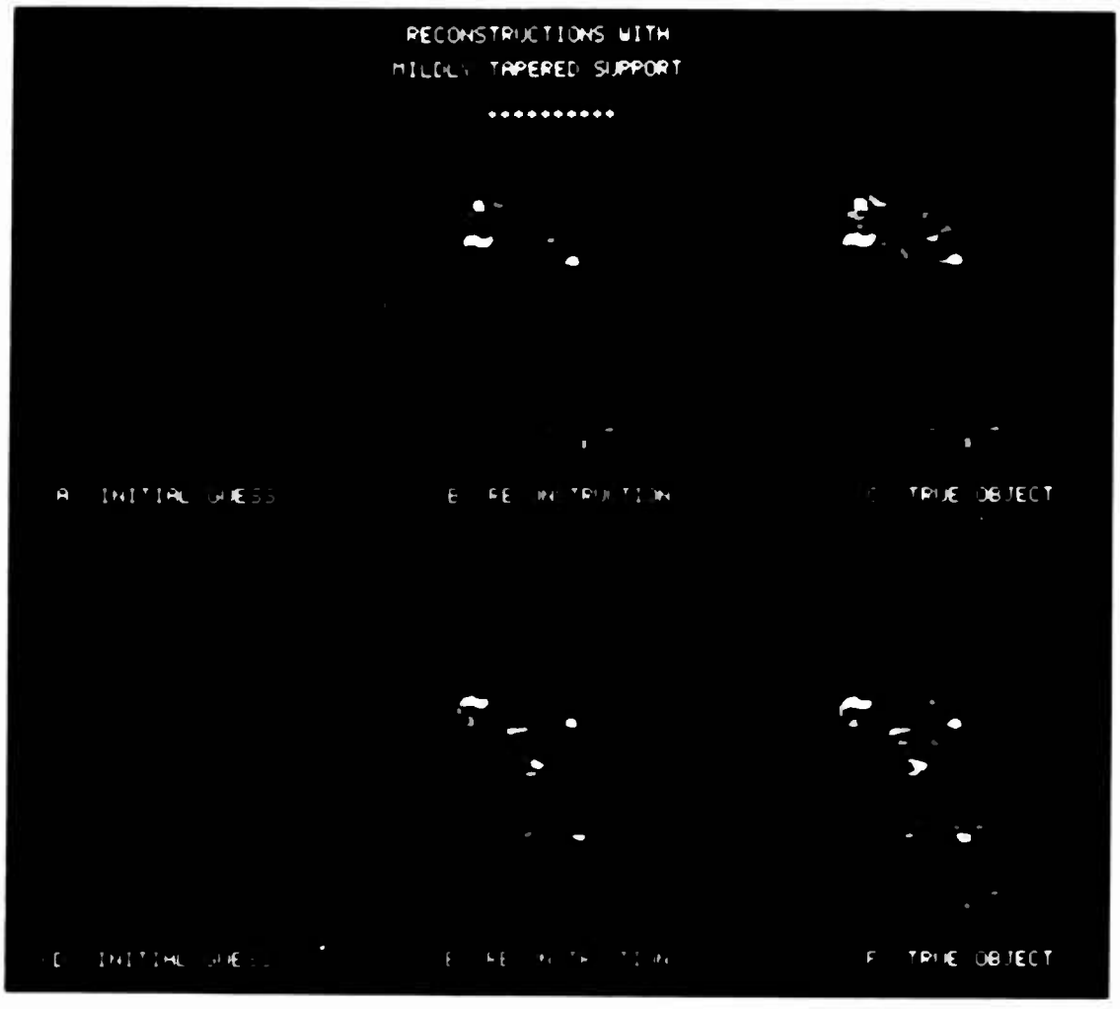



FIGURE 5-6. RECONSTRUCTIONS OF OBJECTS WITH MILDLY TAPERED ILLUMINATION. (Taper #1 in Figure 5-3)

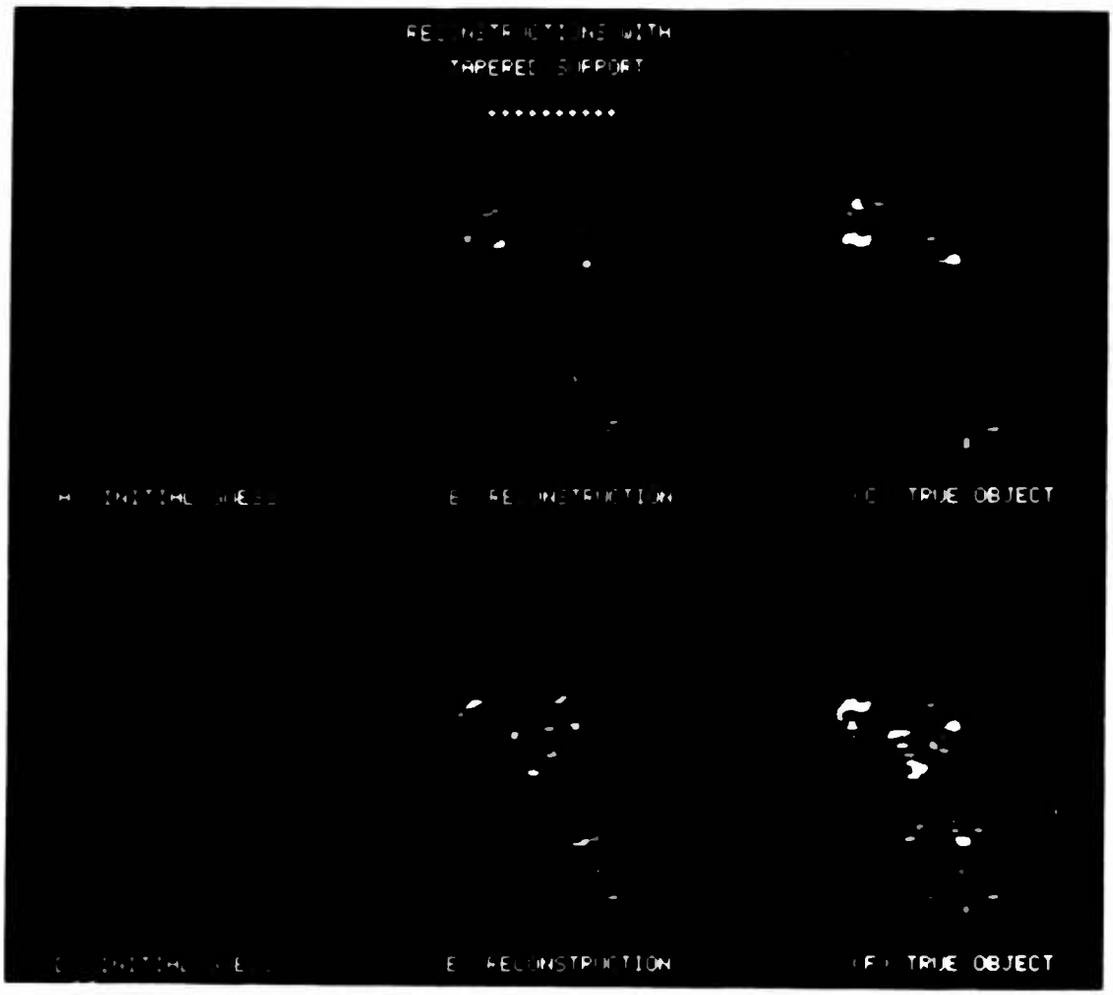



FIGURE 5-7. RECONSTRUCTIONS OF OBJECTS WITH TAPERED ILLUMINATION. (Taper #2 in Figure 5-3)

difference image is bipolar and a bias was added for display in Figure 5-8. Notice that the difference image indicates that the reconstruction is shifted in the horizontal direction relative to the true object. This suggests that the algorithm may be stagnating because of its inability to properly register the reconstruction relative to the support constraint when tapered illumination is used.

To better understand this conjectured mode of stagnation consider an object with tapered illumination $f(x,y)$. We define a binary mask $m(x,y)$ that is the characteristic function of the known support:

$$m(x,y) = \begin{cases} 1, & (x,y)\varepsilon S \\ 0, & (x,y)\varepsilon S' \end{cases} \qquad (5\text{-}6)$$

where S' stands for the complement of S. An image $g'(x,y)$ outputted by the iterative Fourier transform algorithm is the inverse Fourier transform of a Fourier-domain estimate having modulus equal to the given Fourier modulus data coupled with the current estimate of the Fourier phase. Suppose the output image is just a shifted version of the object:

$$g'(x,y) = f(x - x_0, y - y_0) \qquad (5\text{-}7)$$

A shift in the object domain introduces a linear phase factor in the Fourier domain and has no effect on the Fourier modulus. This output image will clearly satisfy the Fourier modulus constraint. The output image has, however, been shifted relative to the mask so that the object domain support constraint has been violated. In other words, multiplying by the mask function will crop an edge of the output image. We use a normalized error metric to indicate the degree of inconsistency between an estimate and the object support constraint:

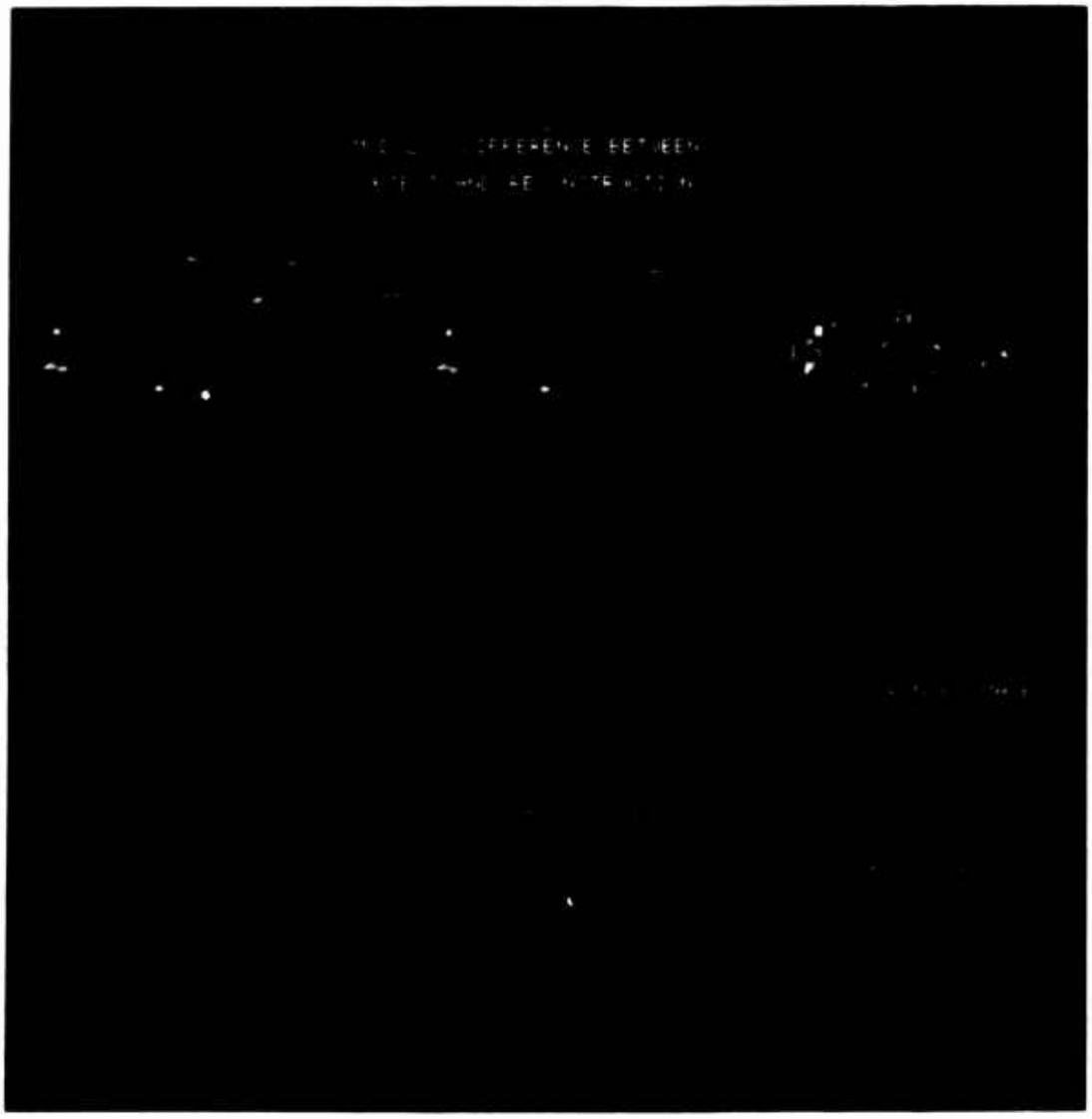


FIGURE 5-8. MODULUS DIFFERENCE BETWEEN OBJECT AND RECONSTRUCTION. (Illumination due to Taper #2 in Figure 5-3) The bipolar difference image has been biased up for display.

$$E_o^2 = \frac{\sum_{x,y} |g'(x,y)m'(x,y)|^2}{\sum_{x,y} |g'(x,y)|^2} \qquad (5\text{-}8)$$

where $m'(x,y)$ is the characteristic function of $S'$. If the shift vector $(x_o,y_o)$ is small with respect to the illumination taper the object domain error metric will also be relatively small. This is because only the tapered edges, where there is little energy, will be cropped and this contributes to only a small portion of the total object energy.

Though the cropped output image now satisfies the support constraint, its Fourier-transform modulus no longer exactly equals $|F(u,v)|$. It can easily be shown that the Fourier-domain error metric is also small. Thus the error metric penalty is small in either domain when shifting a tapered object by a small amount. An algorithm that chooses successive estimates based upon these error metric objective functions will be insensitive to small shifts and would easily stagnate due to extremely small slopes in the objective function. Such an algorithm would be ineffective at finding the proper object registration. Furthermore, one can imagine that with the right redistribution of the cropped object energy an object estimate could correspond to a local minimum in the objective function.

Although the mode of stagnation just presented is conjecture, it provides the motivation for the "shrunken-mask" algorithm. The shrunken-mask algorithm is designed to find the proper registration early on in the iterative reconstruction thus circumventing shift-related stagnation that might otherwise appear.

Consider a new binary mask $m_t(x,y)$ created by hardlimiting the tapered illumination function with some intermediate threshold value:

$$m_t(x,y) = \begin{cases} 1, & (x,y) \text{ such that } |w(x,y)| > t \\ 0, & (x,y) \text{ such that } |w(x,y)| \leq t \end{cases} \tag{5-9}$$

where t is the threshold value, $0 \leq t \leq 1$. Notice that $m_t(x,y)$ will be a "shrunken" version of the full mask m(x,y) defined for t = 0. Suppose that we employ the shrunken mask as the support constraint. If we crop the true object with the shrunken mask this will yield an estimate with a modest penalty in both the object and Fourier domains, so long as the threshold value is not too large. Notice, however, that a shift in this cropped estimate will yield an object-domain penalty, due to the shrunken mask and the artifically created discontinuous object edges, that is much greater than the penalty that would be due to the normal support constraint. Thus one would expect the output image to be centered better with the shrunken mask.

While the Fourier modulus and the shrunken-mask support constraints are inconsistent, they may still be jointly enforced in an iterative reconstruction algorithm to get an intermediate reconstruction. We might expect this intermediate result to display gross features of the true object in proper registration. Enlarging the mask to its full size (setting t = 0) removes the constraint inconsistency and allows for a complete reconstruction that hopefully avoids shift-related local minima. The shrunken-mask algorithm is shown schematically in Figure 5-9.

The shrunken-mask algorithm was first tested on the elliptical object where the taper (taper #2 in Figure 5-3) induced stagnation in previous trials. The convergence characteristics are displayed in Figure 5-10. It is clear that the conventional algoritnm performed better early on in the iterative sequence. This is reasonable since the support constraint is initially looser and easier to satisfy. By contrast, the shrunken-mask algorithm error metric quickly levels off while an intermediate reconstruction is being produced but drops dramatically when the full-size mask is introduced.

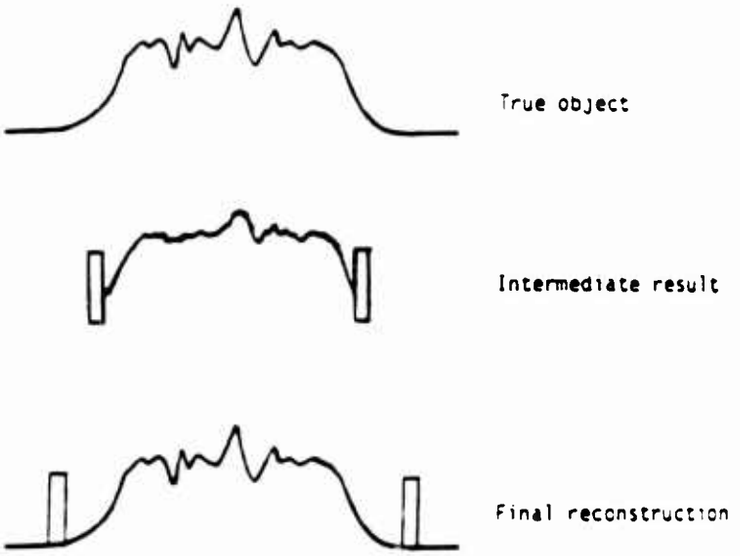



**FIGURE 5-9. THE SHRUNKEN-MASK ALGORITHM**

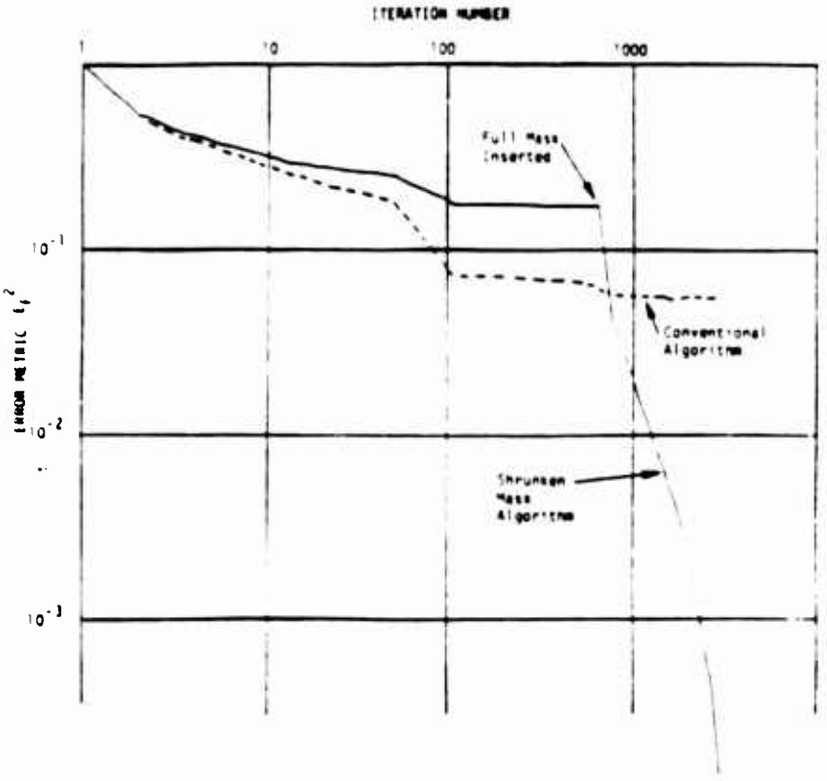



**FIGURE 5-10. CONVERGENCE FOR SHRUNKEN-MASK ALGORITHM. Taper is Taper #2 in Figure 5-3. The shrunken mask had a threshold value t = .9.**

## 5.4 THE ENLARGING MASK ALGORITHM

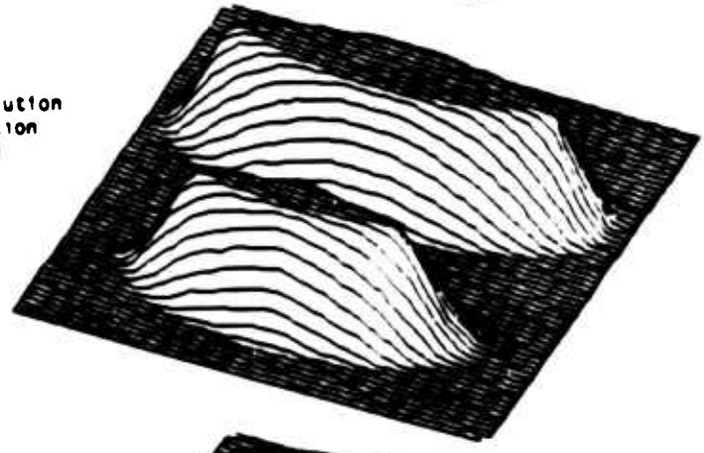
While the success in the shrunken-mask algorithm is encouraging the amount of illumination taper for which it worked remains extremely small. A much more substantial taper was introduced by using a circular convolution kernel with a radius of 4 pixels. The resultant illumination pattern is shown in Figure 5-11. When the shrunken-mask algorithm was applied to an object with this illumination the convergence was not much better than the conventional algorithm. This was true for a variety of threshold values that were tested. Apparently the increased taper is a significant obstacle for the shrunken-mask algorithm.

Recall that the shrunken-mask algorithm jumps from a small mask to the full mask in a single step. A logical generalization of the shrunken-mask algorithm uses several intermediate-size masks in order to make a more gradual transition to the full size mask. We call this the "enlarging-mask" algorithm. The collection of masks used in a given application is characterized by a sequence of threshold values. The convergence curve for the enlarging-mask algorithm when applied to an object with this increased taper is shown in Figure 5-12. The scallop effect exhibited by the convergence curve is due to the successive application of increasingly enlarged masks. The enlarging-mask algorithm clearly out-performs the shrunken-mask algorithm and the final reconstruction exhibits very good agreement with the data and support constraint.

A final trial was performed with an even more realistic illumination taper created with a Gaussian-like convolution kernel with a maximum radius of 6 pixels. This kernel, K(r), was formed by correlating a circle function with a radius of 2 pixels with its own autocorrelation:

A. No taper

B. Taper due to convolution with a circle function (radius = 4 pixels)



C. Taper due to convolution with a Gaussian-like function (max radius = 6 pixels)

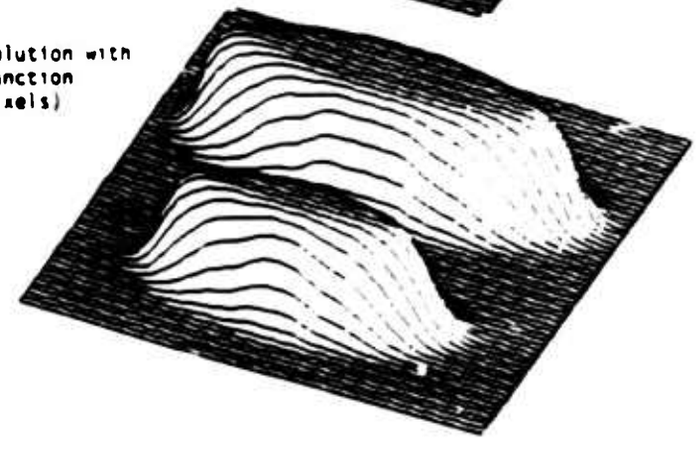


FIGURE 5-11. ILLUMINATION PATTERNS

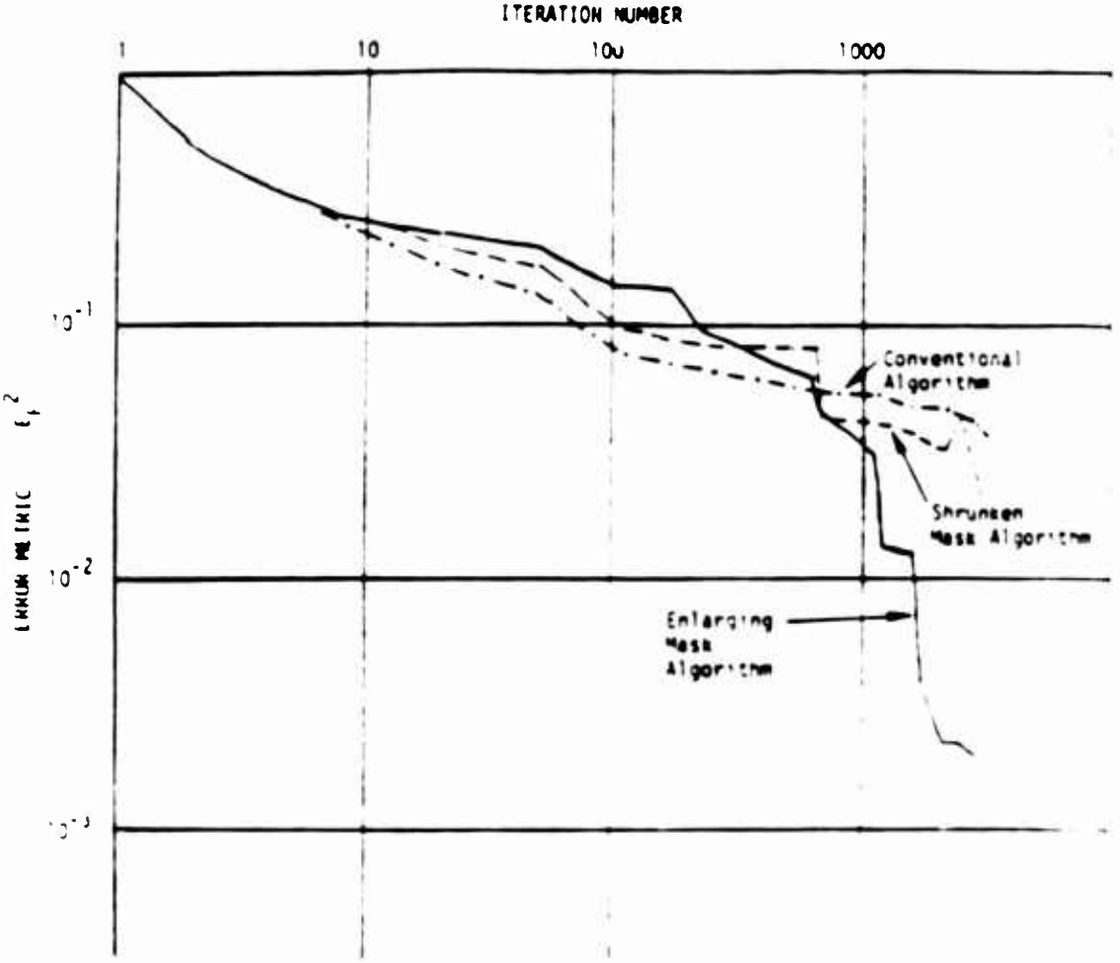



FIGURE 5-12. CONVERGENCE BEHAVIOR FOR ALL THREE ALGORITHMS
Illumination taper is due to a circular convolution kernel (radius = 3 pixels). Threshold sequence for shrunken-mask algorithm: .5, .3, .2, .10

$$K(r) = \text{CIRC}\ (r/2) ** \text{CIRC}\ (r/2) ** \text{CIRC}\ (r/2), \qquad (5\text{-}10)$$

where the double star indicates two-dimensional crosscorrelation. This kernel is a close approximation to a two-dimensional Gaussian function. The resultant illumination pattern is shown in Figure 5-11. Note that this illumination has a smoother taper and that the tails extend out further at very low levels. The convergence curves for this case are shown in Figure 5-13. Again the enlarging-mask algorithm succeeds at finding a reconstruction that is in excellent agreement with the data and support constraint whereas the conventional algorithm did not. This reconstruction is visibly indistinguishable from the true object. The results of reconstructions performed with and without the enlarging-mask algorithm are given in Figure 5-14 for illumination patterns due to the circular and Gaussian convolution kernels.

While tapered illumination presents significant stagnation problems for conventional phase-retrieval algorithms, these examples demonstrate that the enlarging-mask algorithm successfully circumvents these difficulties, even in the presence of large amounts of taper.

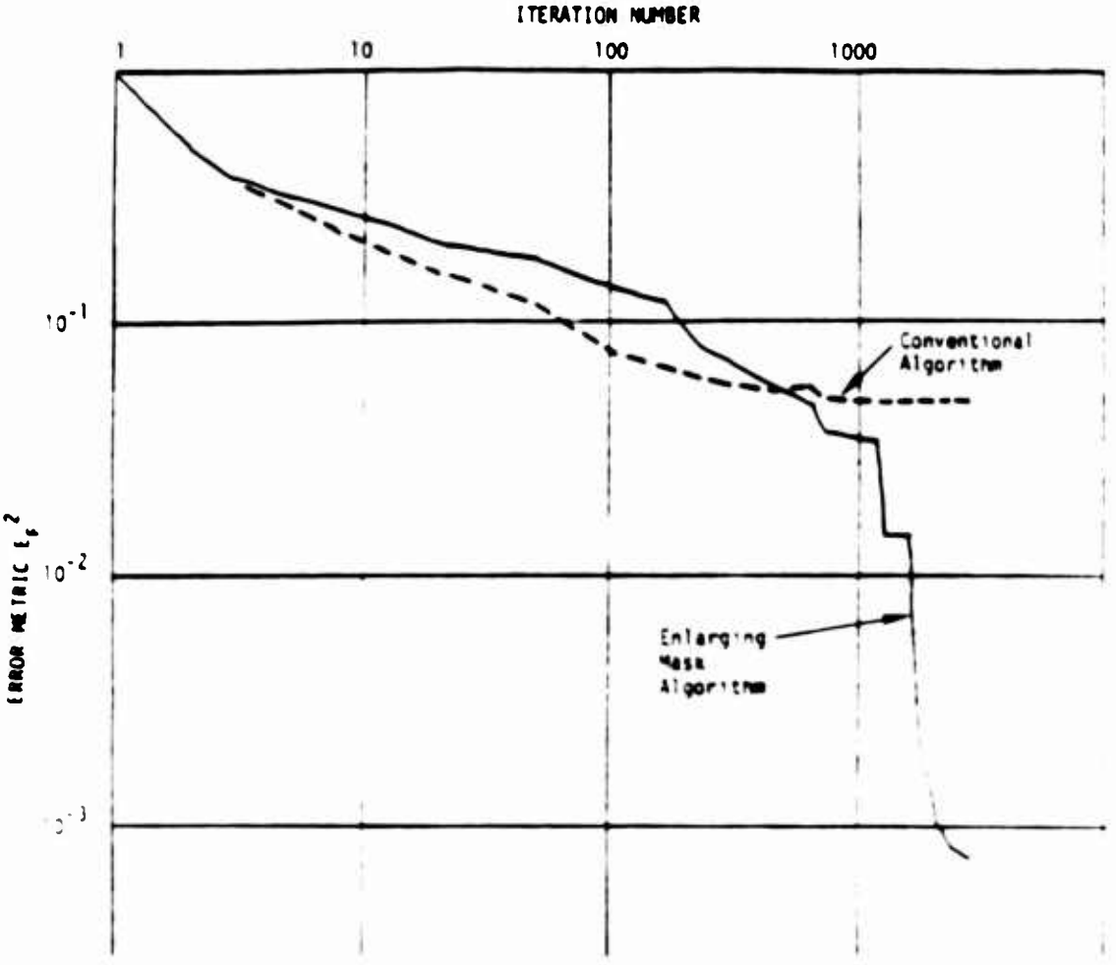



FIGURE 5-13. CONVERGENCE BEHAVIOR WHEN THE TAPER IS DUE TO A GAUSSIAN-LIKE CONVOLUTION KERNEL (max radius = 6 pixels). Threshold sequence for enlarging-mask algorithm: .5, .3, .2, .1, .001.

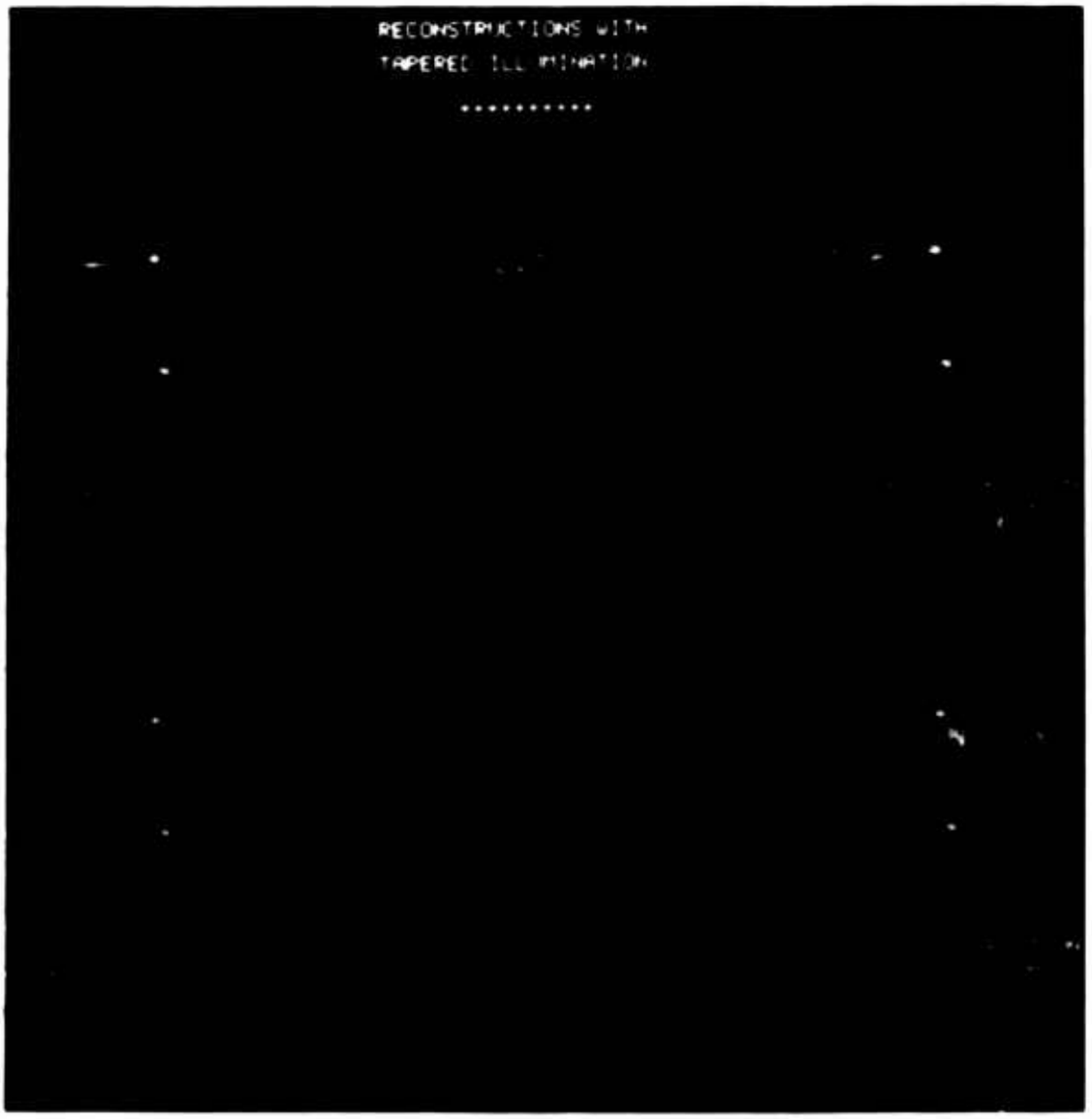



FIGURE 5-14. RECONSTRUCTIONS WITH AND WITHOUT THE ENLARGING-MASK ALGORITHM (EMA). The illumination pattern in A-C is shown in Figure 5-11B. The illumination pattern in D-F is shown in Figure 5-11C.

## References

5.1. J.R. Fienup, "Phase Retrieval from a Single Intensity Distribution," in Optics in Modern Science and Technology, Conference Digest for ICO-13, 20-24 August 1984, Sapporo, Japan, pp. 606-609.

5.2. J.R. Fienup, "Phase Retrieval Using a Support Constraint," IEEE ASSP Workshop on Multidimensional Digital Signal Processing," Leesburg, VA, 28-30 October 1985.

5.3. J.R. Fienup, "Phase Retrieval Algorithms: A Comparison," Appl. Opt. 21, 2758-2789 (1982).

# 6
# GRADIENT-SEARCH METHODS IN PHASE RETRIEVAL

## 6.1 INTRODUCTION

Researchers have explored many approaches to solving the phase retrieval problem. These include direct methods using complex zeros in the analytically extended Fourier modulus [6.1], the error-reduction algorithm [6.2, 6.3], input-output algorithms [6.3], recursive algorithms [6.4, 6.5], and gradient-search algorithms [6.3, 6.6, 6.7, 6.8]. Of these approaches the input-output algorithms or more specifically the hybrid input-output (HIO) algorithm appears to be the current algorithm of choice when operating on 2-dimensional data. The HIO algorithm has consistently outperformed competing algorithms with respect to computational burden and robustness to noise. In spite of the relative success of the input-output algorithms there are documented instances in which such an algorithm converges extremely slowly or even stagnates in its convergence [6.9].

In this report we are interested in the specific phase-retrieval problem for which the Fourier modulus and an object support constraint are known. We resurrect the idea of employing a gradient-search method in the hopes of developing an algorithm that will compete well with or complement the input-output approach. Gradient-search approaches require the determination of an objective function that indicates the degree of consistency with the data and the constraints. This choice is pivotal in designing a specific gradient-search algorithm. We propose here three distinct objective functions and explore the performance of each when used in conjunction with standard gradient-search techniques. In the next section we discuss the error-reduction algorithm, the parent of the input-output algorithms and indicate how it can be interpreted as a gradient-search algorithm. We introduce the first new objective function, called the summed objective function, in Section 6.3. The second and third objective functions are introduced in Sections 6.4 and 6.5. These objective functions utilize the same object-support error

metric but differ in their underlying parameters. We conclude in Section 6.6 with projections of future work.

## 6.2 THE ERROR-REDUCTION ALGORITHM

An iterative algorithm that has enjoyed much success in phase retrieval is known as the error-reduction (ER) algorithm, which may be easily understood by referring to Figure 6-1. This algorithm consists of transforming between object and Fourier domains and applying appropriate constraints in the respective domains. We use the symbol $g_k(x)$ to represent the estimate of a object given by the kth iteration of the ER algorithm. The prime notation in $g_k'(x)$ indicates a version of the kth estimate for which the Fourier-domain constraints have been enforced. We use uppercase symbols to denote a Fourier-domain representation of a function. In practice the data are always sampled and therefore we use the discrete Fourier transform (DFT)

$$G(u) = \sum_x g(x) e^{-i2\pi u \cdot x/N} \qquad (6\text{-}1)$$

and its inverse

$$g(x) = N^{-2} \sum_u G(u) e^{i2\pi u \cdot x/N} \qquad (6\text{-}2)$$

in the algorithm. Of course the DFT is most efficiently computed with a fast Fourier transform (FFT). In Eqs. (6-1) and (6-2) x and u are two-dimensional vectors in the object and Fourier domains, respectively, and the summation notation is understood to represent a separate summation for each component of the vector running from 0 to N-1.

In order to enforce a given constraint we define a least-squares constraint operator. The function of the operator is to produce an output that conforms to the constraint but differs from the input as little as possible in a least-squares sense. It can be shown that when the constraint operators have this property the mean squared error

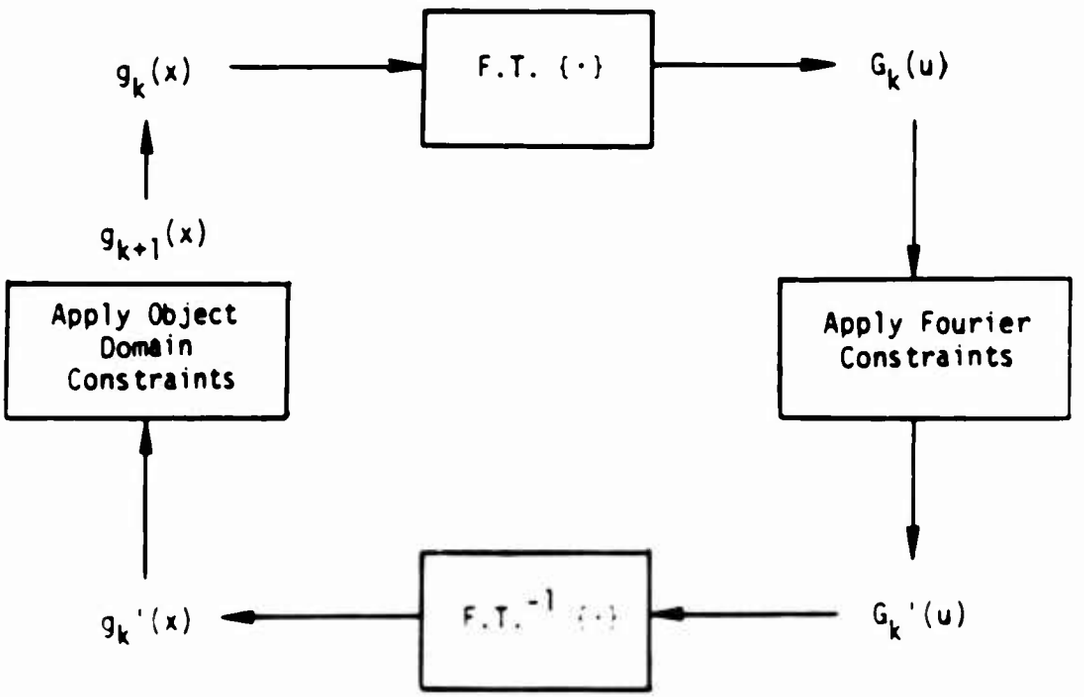



FIGURE 6-1. ERROR REDUCTION ALGORITHM

between the latest estimate and the data or known information will decrease (or stay the same) at each iteration [6.3]. Thus as the algorithm proceeds, the reconstruction estimate conforms more and more closely to the given constraints. This is the motivation for the title "error reduction."

Although many types of constraints have been used with the ER algorithm, our problem affords a modulus constraint in the Fourier domain in conjunction with a support constraint in the object domain. This specific realization of the ER algorithm is illustrated in Figure 6-2. The modulus constraint is performed by substituting the modulus of the latest estimate with the known modulus while leaving the Fourier phase untouched:

$$G'(u) = \frac{G(u)|F(u)|}{|G(u)|} \qquad (6\text{-}3)$$

where $F(u)$ is the known Fourier modulus. The object domain constraint is equally straightforward and is enforced by setting the values of all pixels that fall outside of the support equal to zero:

$$g_{k+1}(x) = \begin{cases} g_k'(x), & x \in S \\ 0, & x \in S' \end{cases} \qquad (6\text{-}4)$$

where S stands for the set of pixels within the known support and S' is the complement of S.

In order to monitor the progress of the ER algorithm it is useful to define an error metric for each of the constraints. The error metric is essentially a mean squared error between estimates before and after a constraint has been applied and indicates the degree of agreement between the latest estimate and the known constraint. The error metric for the Fourier modulus constraint is defined as follows:

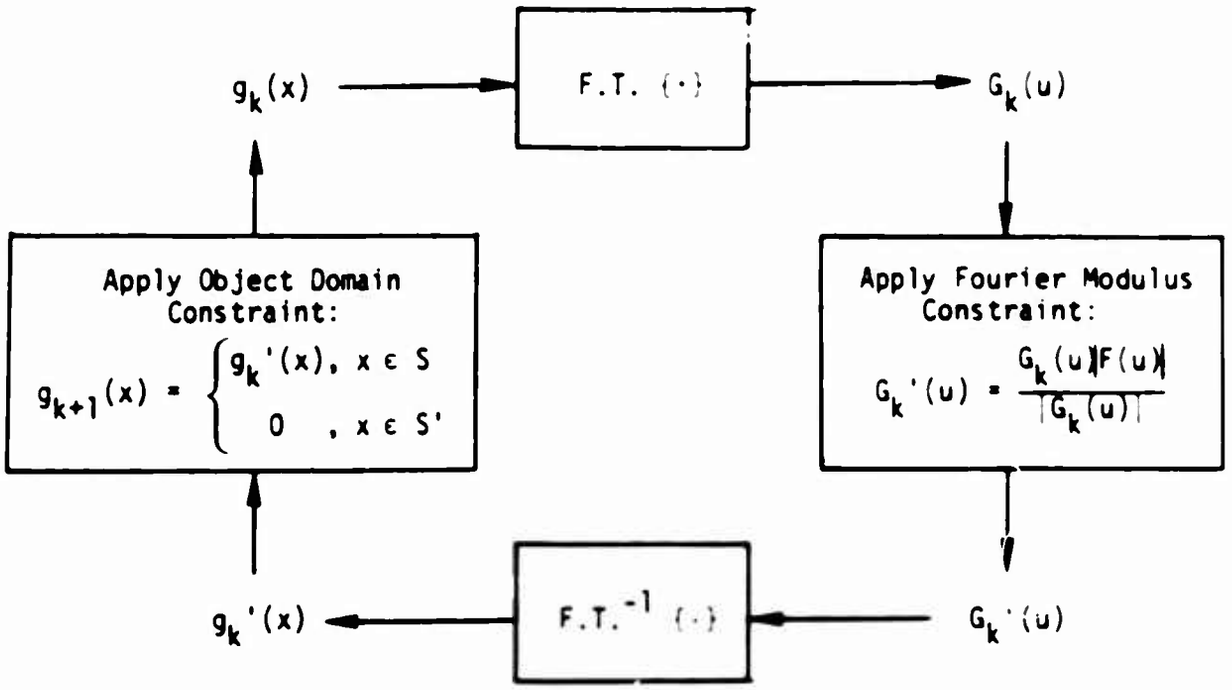



FIGURE 6-2. ERROR REDUCTION ALGORITHM FOR FOURIER MODULUS AND OBJECT SUPPORT CONSTRAINTS

$$e_F^2 = N^{-2} \sum_u [|G(u)| - |F(u)|]^2 \quad . \qquad (6\text{-}5)$$

The error metric for the object-support constraint is given by

$$e_0^2 = \sum_{x \epsilon S'} \left| g'(x) \right|^2 \quad . \qquad (6\text{-}6)$$

As the algorithm proceeds both of these error metrics will decrease. If they simultaneously achieve values close to or equal to zero then the algorithm has achieved a restoration that has good agreement with both constraints.

Suppose that we treat the error metric $e_F^2$ as an objective function to be used in a gradient-search algorithm. Our desire is to minimize the objective function by varying a set of parameters in the estimate. The parameters we employ are the individual pixel values of the estimate. For the present we treat only real-valued objects which require $N^2$ independent parameters for an NxN image (complex objects require $2N^2$ parameters). The $j^{th}$ pixel in the object domain is located by a vector $x_j$ where the subscript j represents any convenient ordering of the $N^2$ pixels. We construct an $N^2$-dimensional Euclidian vector space for which each coordinate axis corresponds to an individual parameter. Each point in this parameter space therefore corresponds to an object estimate and may be represented by the parameter vector g(x). We represent the $j^{th}$ parameter and its associated parameter space unit vector by $g(x_j)$ and $v_j$, respectively. The unit vector $v_j$ may be interpreted as an estimate for which all pixels are zero except for the $j^{th}$ pixel which has unit strength. This vector may also be represented by the Kronecker delta $\delta_{x,x_j}$. The objective function, $e_F^2(g(x))$, is a function of the $N^2$ parameters, and may be visualized as a surface in an $N^2+1$ -dimensional space. If we were able to calculate the gradient of this surface at given estimate locations then well-known gradient-search methods could be employed. The gradient is formally expressed

$$\nabla e_F^2(g(x)) = \sum_{j=1}^{N^2} \frac{\partial e_F^{\,2}}{\partial g(x_j)} \, v_j \quad . \tag{6-7}$$

One method of computing the gradient is to proceed numerically using a finite differences approximation to the partial derivative:

$$\frac{\partial e_F^{\,2}}{\partial g(x_j)} \cong \frac{e_F^{\,2}(g(x) + \alpha v_j) - e_F^{\,2}(g(x))}{\alpha} \tag{6-8}$$

where $\alpha$ is small compared with significant feature sizes in the objective surface. This brute-force approach is computationally prohibitive since each evaluation of $e_F^{\,2}$ involves an NxN FFT and this must be accomplished for each of the $N^2$ parameters. Fortunately Fienup [6.3] showed that the exact partial derivative may be calculated analytically as follows:

$$\frac{\partial e_F^{\,2}}{\partial g(x_j)} = 2\left[g(x_j) - g'(x_j)\right] \quad . \tag{6-9}$$

We reemphasize that the prime indicates that the Fourier magnitude constraint has been applied to the estimate. If Eq. (6-9) is substituted into Eq. (6-7) the result implies that the entire gradient may be evaluated with a forward and an inverse FFT:

$$\begin{aligned} \nabla e_F^2(g(x)) &= \sum_j 2\left[g(x_j) - g'(x_j)\right] v_j \\ &= 2\left[g(x) - g'(x)\right] \end{aligned} \quad . \tag{6-10}$$

This desirable result means that a gradient search method could realistically be employed for the $e_F^{\,2}(g(x))$ objective function.

Perhaps the simplest gradient-search algorithm is the method of steepest descent [6.10]. According to this approach the latest estimate may be improved upon by moving in parameter space in a direction opposite that of the gradient. The location of the minimum of the objective function along the resulting one-dimensional cut is then determined giving an improved estimate. This procedure is repeated iteratively until a local minimum in the objective function is achieved.

Some optimization problems afford additional a priori information about disallowed regions in parameter space. There are many ways of constraining the final solution to the allowed region of parameter space. One obvious way of incorporating this information is to proceed as usual with the steepest-descent algorithm until an estimate is produced that violates the a priori knowledge. A constraint operator is then employed to find the closest allowed estimate. The steepest-descent algorithm is then applied to the latest allowed estimate. Unfortunately this constrained steepest-descent algorithm can be very slow since the direction of steepest descent is often in competition with the direction enforced by the constraint operator.

A careful analysis of the ER algorithm reveals that it is, in fact, a constrained steepest-descent algorithm for which the objective function is $e_F^2(g(x))$ and the knowledge of object support defines a disallowed region in parameter space. The kth iteration of the ER algorithm begins with an estimate, $g_k(x)$, and replaces its Fourier modulus with the known Fourier modulus to get $g_k'(x)$. Notice that this intermediate result is equivalent to moving from $g_k(x)$ in parameter space in a $g_k'(x) - g_k(x)$ direction; that is, in a direction opposite to that of the gradient. In fact it can be shown that the objective function is a minimum (zero) at $g_k'(x)$. Typically $g_k'(x)$ will violate the known support and therefore exists in a disallowed region in parameter space. Applying the support constraint to $g_k'(x)$ produces a

new estimate, $g_{k+1}$, that now resides in the allowed region, thus completing one iteration of the constrained steepest-descent algorithm.

While we have thus far treated the Fourier-domain error metric as an objective function we could just as easily have selected the object-domain error metric, $e_o^2(g'(x))$, for that role. The gradient for this objective function is easily obtained because the calculation of the partial derivative with respect to a pixel value is more direct:

$$\frac{\partial e_o^2}{\partial g'(x_j)} = \frac{\partial}{\partial g'(x_j)} \sum_{x \in S'} [g'(x)]^2 \qquad (6\text{-}11)$$

$$= \begin{cases} 0 & x_j \in S \\ 2g'(x_j) & x_j \in S' \end{cases} .$$

Recall that the support constraint operator sets to zero all pixels in S' and leaves those in S untouched. Clearly, this operation moves the latest estimate $g_k'(x)$ in a direction opposite that of $\nabla e_o^2(g'(x))$. In addition this objective function is quadratic along this one-dimensional cut with a minimum value (zero) at $g_{k+1}(x)$. The Fourier modulus constraint may now be interpreted as the operator that takes $g_{k+1}(x)$ out of a new disallowed region in parameter space. Thus the ER algorithm qualifies as a constrained steepest-descent algorithm from this new perspective as well.

## 6.3 THE SUMMED OBJECTIVE FUNCTION

Historically the error metric $e_o^2$ has been used to evaluate an estimate for which the Fourier constraints have been satisfied. Consequently, this error metric is a function of the pixel values in a primed estimate, as defined in Eq. (6-6). A simple generalization of this definition yields a new error metric that can be applied to any estimate g(x):

$$\varepsilon_0^2(g(x)) = \sum_{x \in S'} [g(x)]^2 \qquad (6\text{-}12)$$

It is easy to show that the partial derivative of $\varepsilon_0^2$ with respect to pixel values in the estimate has the same form as given in Eq. (6-11). Clearly this generalized objective function and its gradient still pertain to functions for which the Fourier constraints have been satisfied. Notice, however, that $\varepsilon_0^2(g(x))$ now has the same underlying parameters as $e_F^2(g(x))$. This observation affords still a third interpretation of the ER algorithm that yields new insight. The ER algorithm may be viewed as alternately performing steepest-descent operations on two objective functions, $e_F^2(g(x))$ and $\varepsilon_0^2(g(x))$, that coexist in the same parameter space. In practice it is often observed that the ER algorithm converges rapidly for iterations early in the sequence but that convergence becomes painfully slow as the iteration number increases. This is because the work performed in minimizing the $e_F^2(g(x))$ objective function is largely nullified when minimizing the $\varepsilon_0^2(g(x))$ objective function, and vice versa. Figure 6-3a illustrates this point pictorially. This viewpoint suggests the definition of a new objective function that is the sum of the opposing objective functions:

$$e_s^2(g(x)) = e_F^2(g(x)) + \varepsilon_0^2(g(x)) \; . \qquad (6\text{-}13)$$

The gradient of this new objective function is simply the sum of the gradients already derived:

$$\nabla e_s^2(g(x)) = \nabla e_F^2(g(x)) + \nabla \varepsilon_0^2(g(x)) \; . \qquad (6\text{-}14)$$

The calculation of this new gradient involves a forward and inverse FFT and a small amount of computational overhead. Figure 6-3b suggests how moving in a direction opposite that of the gradient of the summed

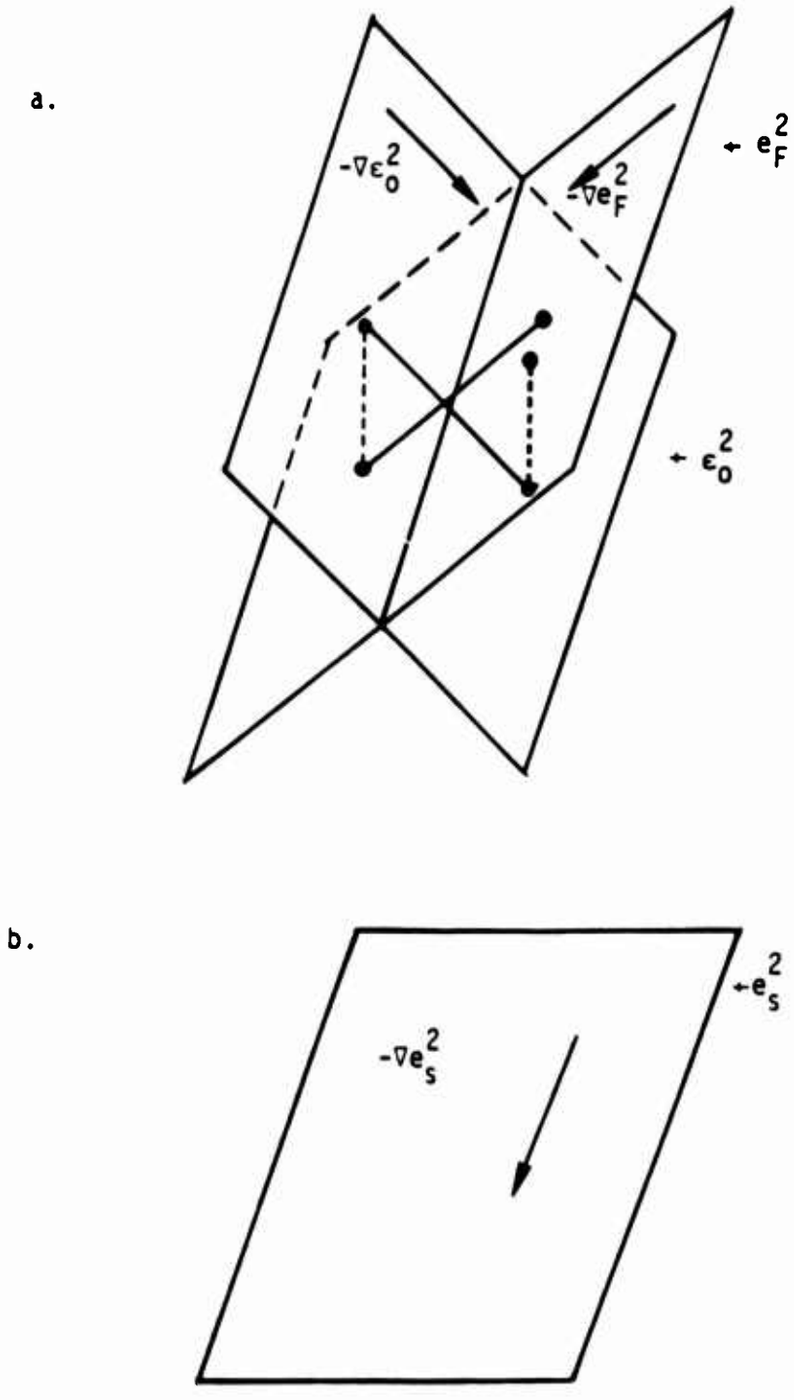



FIGURE 6-3. OBJECTIVE FUNCTION SURFACES FOR TWO PARAMETER OBJECTS
a. Surfaces used in error reduction. b. Summed objective function surface.

objective function may circumvent stagnation due to opposing constraints. Notice that if we choose to remain with steepest descent using $e_s^2$, the stepsize still has to be determined. This can be accomplished by one of a variety of line search methods that utilize additional samples of the objective function. Each additional objective-function evaluation requires a single forward FFT. Furthermore, because the gradient of the summed objective function is so easily computed more sophisticated gradient-search methods such as the method of conjugate gradients or a memoryless quasi-Newton method [6.10] may profitably be employed. Finally, a simple generalization of these ideas to include complex objects is found in Appendix E.

## 6.4 THE $e_o^2(g(x))$ OBJECTIVE FUNCTION

We now briefly review the basic characteristics of the so-called input-output phase-retrieval algorithms. These observations will suggest the defining of a new objective function that will serve as an alternative to the summed objective function.

It is convenient to partition an iteration of the ER algorithm into two steps. The first step enforces the Fourier-domain constraints while the second step enforces the object-domain constraints. For the moment we focus on the first step. This step involves a Fourier transformation of the latest estimate, a substitution of the Fourier modulus by the known values, and an inverse Fourier transformation. Together, these operations constitute the enforcement of Fourier knowledge and may be viewed as a single nonlinear operation. This is depicted schematically in Figure 6-4. It is important to recognize that any output of this operation will satisfy the Fourier-domain constraints and consequently $e_F^2$ will be zero. Should the output also satisfy the object-domain constraints then a solution has been found. This suggests that clever adjustments to the input function might produce an output that more closely satisfies the object-domain constraints. The degree of consistency with the support constraint can be monitored by the $e_o^2$

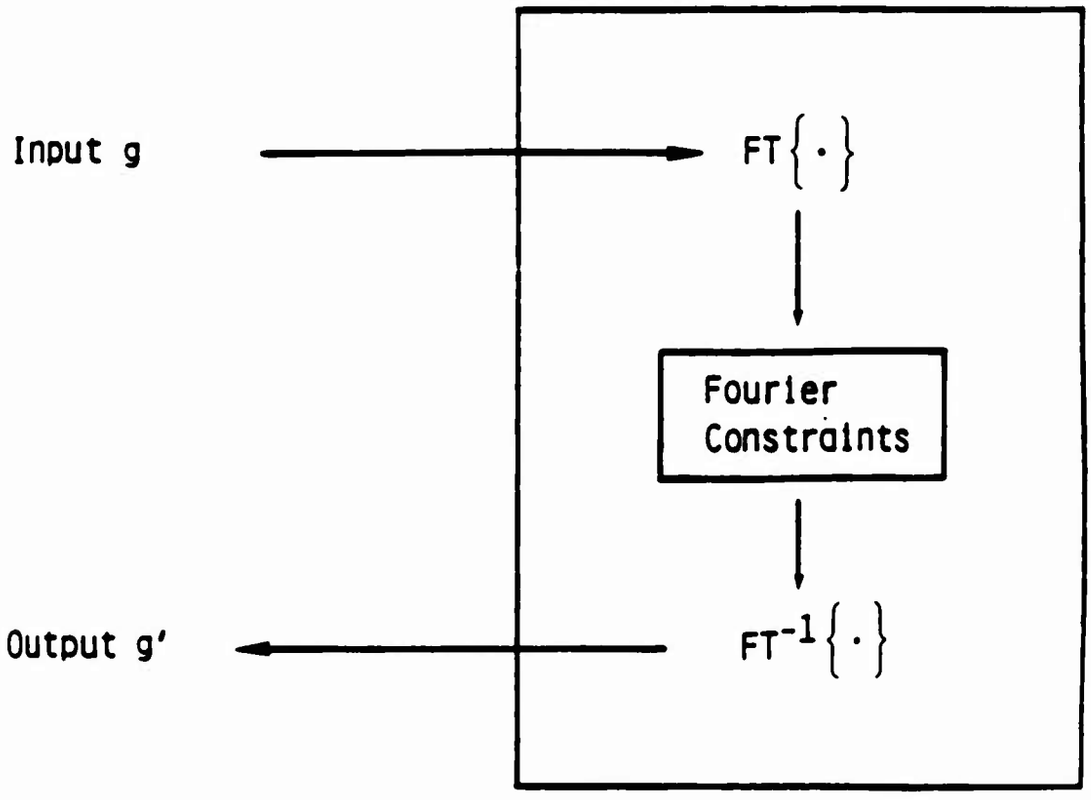



FIGURE 6-4. INPUT-OUTPUT ALGORITHM

error metric defined in Eq. (6-6). A variety of feedback strategies borrowed from nonlinear-systems control theory can be employed to modify the latest input in order to drive the $e_0^2$ error metric toward zero. The use of each feedback rule defines an individual algorithm and the collection of feedback rules defines the class of input-output phase-retrieval algorithms. All feedback rules that have been employed to date are point operations meaning that an input pixel-value adjustment is based solely upon the desired change in the corresponding output pixel value.

We recognize immediately that if a solution were to serve as an input function then it will pass through the nonlinear modulus operator unchanged. Notice however that other inputs can also output a solution. In fact any input function with the proper Fourier phase will produce a solution. Thus a solution will result from any of an uncountable infinity of input functions, many of which differ dramatically from the solution. The ER algorithm may be viewed as a particular input-output algorithm for which the feedback rule drives the input (as well as the output) toward a solution. By contrast most input-output algorithms have a more flexible feedback rule since they may converge upon any of the many input functions that yield a true solution upon output.

We reiterate that any output for which $e_0^2$ is zero will be a solution. Therefore, the task of simultaneously minimizing the Fourier and object-domain error metrics has been converted into minimizing a single error metric. Unlike the summed objective function, however, this blending of the two error metrics into one is accomplished without resorting to ad-hoc methods such as summing.

We use the term objective function to refer to an error metric in conjunction with a set of underlying parameters. A logical candidate for an alternative objective function suggested by input-output algorithms is the $e_0^2$ error metric as a function of <u>input</u> pixel values. This new objective function should not be confused with the $e_0^2(g'(x))$

objective function used in the ER algorithm which treats the $N^2$ object-estimate (output) pixel values as parameters. By contrast the new objective function, $e_o^2(g(x))$, utilizes the input-function pixel values as parameters associated with the object estimate given upon output. Having made this subtle but critical distinction we may now write an expression for the gradient of the $e_o^2(g(x))$ objective function:

$$\nabla e_o^2(g(x)) = \sum_{j=1}^{N^2} \frac{\partial e_o^2}{\partial g(x_j)} v_j \quad . \tag{6-15}$$

As before a numerical computation of the gradient is overwhelming. It is natural to ask if an analytic expression for the gradient can be derived. While the details of this calculation are outlined in Appendix F, we give the suprisingly simple result here:

$$\frac{\partial e_o^2}{\partial g(x_j)} = \sum_u \left[ \frac{|F(u)|G_e(u)}{|G(u)|} - \frac{G'(u)G_e^*(u)}{G^*(u)} \right] e^{i2\pi u \cdot x_j/N} \tag{6-16}$$

where * denotes complex conjugate and $G_e(u)$ is the Fourier transform of an error image $g_e(x)$, where

$$g_e(x) = S'(x)g'(x) \tag{6-17}$$

and

$$S'(x) = \begin{cases} 1 \ , \ x \in S' \\ 0 \ , \ x \in S \end{cases} \tag{6-18}$$

Three FFT operations are required to compute $G_e$ from g. A very important feature of the analytic partial derivative quoted in Eq. (6-16) is that it has the form of a DFT. The implication is that given the expression within the brackets all partial derivatives needed to compute the gradient are provided by a single DFT. Thus the total computational cost of finding $\nabla e_0^2(g(x))$ for a given input function is four FFTs plus minor overhead. With these manageable computational requirements the $e_0^2(g(x))$ objective function may be minimized via various gradient-search algorithms. Some care must be taken in the evaluation of Eq. (6-16) to avoid division by zero. This problem can be circumvented by adding a small constant to the Fourier magnitude of the input function at those spatial frequencies for which $|G(u)|$ is identically zero.

Notice that, like input-output algorithms, there are many input functions to which a gradient-search algorithm can converge for this objective function. This means that the objective function contains many global minima, each equally acceptable for producing a solution as an output. It is conceivable that this multiplicity of input solutions could yield faster convergence rates than an objective function having only a single global minimum (e.g. the summed objective function).

It is useful to recognize that any gradient-search algorithm used in conjunction with the $e_0^2(g(x))$ objective function may also be interpreted as a particular feedback rule in an input-output algorithm. Unlike other feedback rules, however, this rule is not a point operation. In other words, the gradient-search feedback rule is more flexible than other existing rules since many input pixels may be adjusted in order to effect a desired change in a single output pixel.

Unfortunately, there is no guarantee <u>a priori</u> that the $e_0^2(g(x))$ objective function has a surface contour that lends itself to minimization via gradient search. For example the $e_0^2(g(x))$ surface

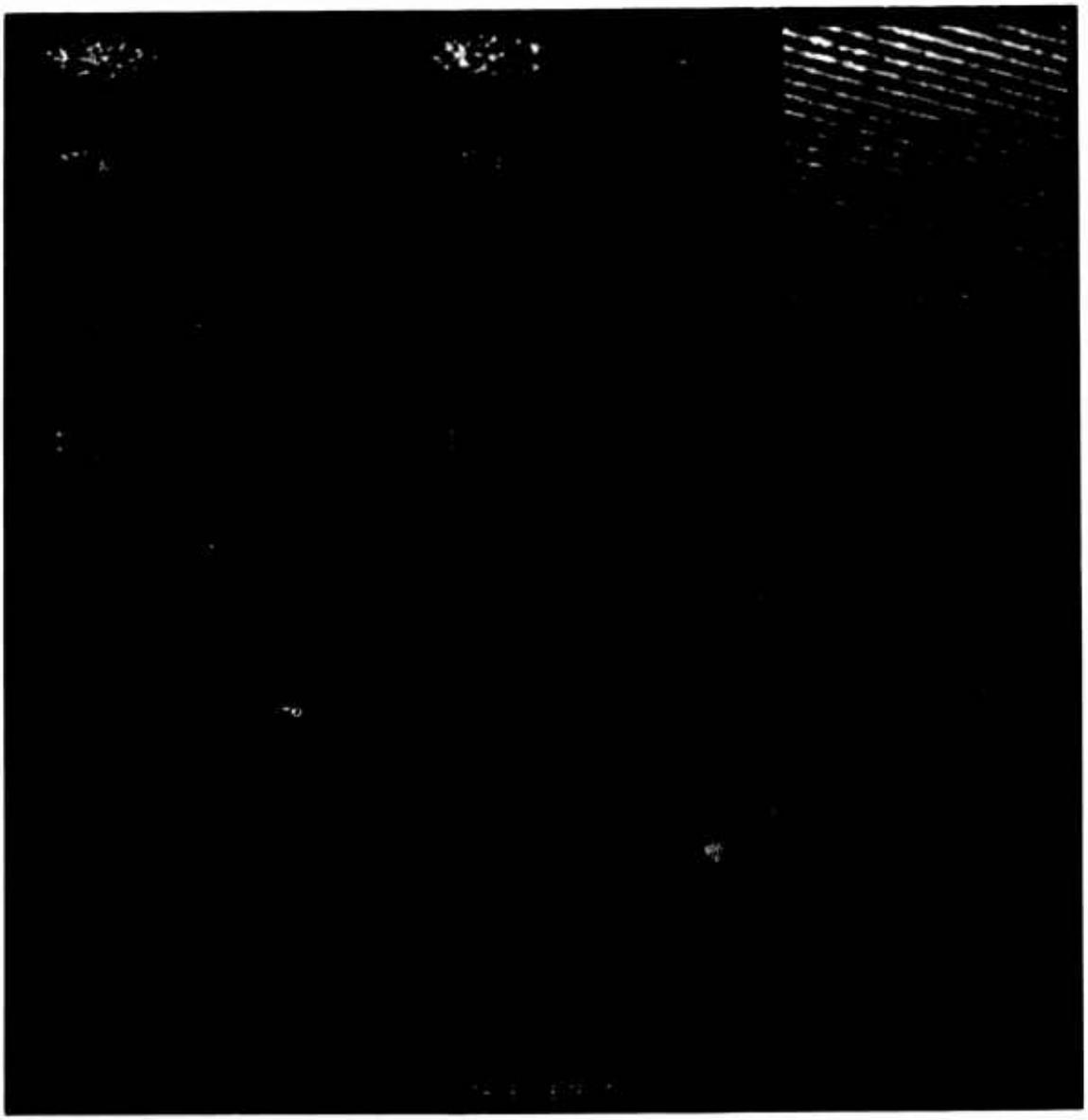


FIGURE 6-5. PRELIMINARY IMAGES DERIVED FROM MINIMIZING THE $e_o^2$ OBJECTIVE FUNCTION

may contain many local minima in which gradient-search algorithms could become entrapped. Answers to such questions are often the byproduct of extensive experimentation.

Some preliminary experiments were performed in which the $e_0^2(g(x))$ objective function was used in conjunction with the method of steepest descent. A number of observations can be made about the results displayed in Figure 6-5. Notice the dominant stripes in the gradient image for the first iteration. By gradient image we mean the image for which each pixel value is assigned the value of the associated component of the gradient. This is the image that is scaled and added to the latest input image to acquire the succeeding input image in a steepest-descent scheme. These stripes are intriguing; but their origin is unknown at present. The magnitude of the gradient was observed to decrease with iteration number. As a result, the stripes from the first gradient image still persist in the 100th input image. Notice, however, that the stripes do not appear in an output image, which is consistent with the notion that the input image need not resemble the output image. It is encouraging that after 100 iterations the output image bears a rough resemblance to the true object. More experimentation with this objective function is needed before a judgement can be made about its usefulness. For example, more sophisticated gradient-search methods would have a better chance of converging to a solution. Should the $e_0^2(g(x))$ objective function in conjunction with the best gradient-search methods prove not to be competitive with current input-output algorithms, it may yet be useful for breaking out of stagnation episodes.

We conclude this section by noting that while we have restricted objects to be real-valued for simplicity, the case admitting complex objects is of great interest when objects are illuminated coherently. The definition and derivation of the gradient of the $e_0^2(g(x))$ objective function for complex objects is presented in Appendix 6.

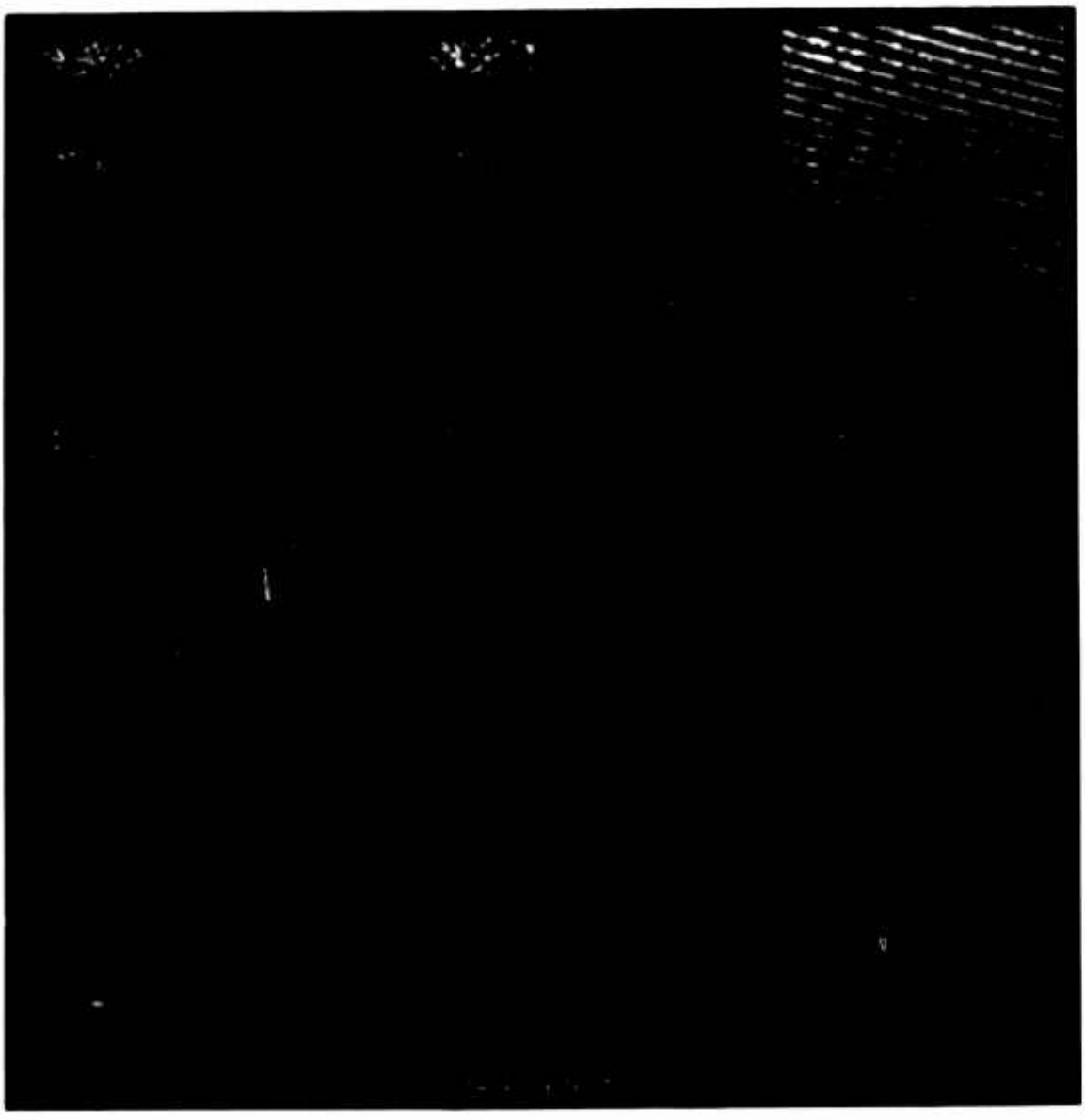


FIGURE 6-5. PRELIMINARY IMAGES DERIVED FROM MINIMIZING THE $e_o^2$ OBJECTIVE FUNCTION

## 6.5 FOURIER PHASE PARAMETERS

The choice of underlying parameters for an objective function can have a tremendous impact upon the behavior of gradient-search algorithms. To this point we have selected the input pixel values (or real and imaginary parts of the input pixels) as our $N^2$ (or $2N^2$) parameters underlying the $e_o^2(g(x))$ objective function. This choice has merit since it affords an analytic expression for the gradient requiring only four FFTs. An alternative and very different set of parameters worth consideration is the set of Fourier phase values in a Fourier estimate of a solution. Because the Fourier modulus is known, a Fourier estimate is determined by an estimate of the Fourier phase, $\phi(u)$:

$$G'(u) = |F(u)|e^{i\phi(u)}. \tag{6-19}$$

An inverse FFT gives the corresponding object-domain estimate,

$$g'(x) = N^{-2} \sum_u |F(u)|e^{i\phi(u)}e^{i2\pi u \cdot x/N} . \tag{6-20}$$

This estimate may also be interpreted as the output from an input-output algorithm since it has the proper Fourier modulus. Consequently, the object-domain error metric can be computed:

$$e_o^2 = \sum_{x \epsilon S'} |g'(x)|^2 . \tag{6-21}$$

The $e_o^2$ error metric is therefore implicitly a function of the Fourier phase values and $e_o^2(\phi(u))$ serves as the third new objective function introduced in this chapter. We mention parenthetically that throughout this section we allow for complex-valued objects since there is no simplification of derivations by resorting to real-valued objects. Notice that the designation of the Fourier phase values as the underlying parameters has fixed the number of parameters at $N^2$. This

is exactly half the number of parameters that occur when using the real and imaginary parts of the input pixel values as parameters. It remains to be seen, though, if an analytic expression for the gradient of the object-domain error metric with respect to the Fourier phase parameters can be derived.

The gradient is defined as

$$\nabla e_o^2(\phi(u)) = \sum_{j=1}^{N^2} \frac{\partial e_o^2}{\partial \phi(u_j)} v_j \tag{6-22}$$

where $v_j$ is the unit vector in parameter space associated with the phase parameter at location $u_j$ in the Fourier-domain estimate. As usual, the heart of the gradient is the partial derivative

$$\frac{\partial e_o^2}{\partial \phi(u_j)} = \frac{\partial}{\partial \phi(u_j)} \sum_{x \in S'} |g'(x)|^2 \tag{6-23}$$

$$= \sum_{x \in S'} g'(x) \frac{\partial g'^*(x)}{\partial \phi(u_j)} + C.C. \tag{6-24}$$

where C.C. stands for complex conjugate. The partial derivative in Eq. (6-24) may be simplified:

$$\frac{\partial g'^*(x)}{\partial \phi(u_j)} = N^{-2} \sum_u |F(u)| e^{-i2\pi u \cdot x/N} \frac{\partial}{\partial \phi(u_j)} e^{-i\phi(u)} \tag{6-25}$$

$$= N^{-2} \sum_u |F(u)| e^{-i2\pi u \cdot x/N} (-i) e^{-i\phi(u)} \delta_{u,u_j} \tag{6-26}$$

Applying the sifting property of the Kronecker delta, $\delta_{u,u_j}$, in Eq. (6-26) leaves only one term from the summation:

$$\frac{\partial g'^*(x)}{\partial \phi(u_j)} = N^{-2} \, |F(u_j)| e^{-i2\pi u_j \cdot x/N} (-i) e^{-i\phi(u_j)} \quad . \tag{6-27}$$

Substituting back into Eq. (6-24):

$$\frac{\partial e_o^2}{\partial \phi(u_j)} = \sum_{x \in S'} \left[ g'(x)(-i)N^{-2} |F(u_j)| e^{-i2\pi u_j \cdot x/N} e^{-i\phi(u_j)} + C.C. \right] \tag{6-28}$$

$$= N^{-2} \, |F(u_j)| \left[ \left( (-i) e^{-i\phi(u_j)} \sum_x S'(x) g'(x) e^{-i2\pi u_j \cdot x/N} \right) + C.C. \right] . \tag{6-29}$$

The summation in Eq. (6-29) is the Fourier error image, $G_e(u)$, defined in the previous section by Eq. (6-17). Therefore we have

$$\frac{\partial e_o^2}{\partial \phi(u_j)} = N^{-2} \, |F(u_j)| \left[ (-i) G_e(u_j) e^{-i\phi(u_j)} + C.C. \right] \tag{6-30}$$

$$= 2N^{-2} \, |F(u_j)| \, \mathrm{Im} \left\{ G_e(u_j) e^{-i\phi(u_j)} \right\} \tag{6-31}$$

where Im $\{\cdot\}$ stands for the imaginary part.

Again we have been able to find an expression for the gradient with a remarkably compact form. Equation (6-31) implies that the component of the gradient associated with the spatial frequency $u_j$ is proportional to the modulus at that spatial frequency and is dependent upon the Fourier-domain error image and the latest Fourier phase in a less direct way. An examination of Eq. (6-31) reveals that the entire gradient can be computed with 2 FFTs plus minor overhead. The actual evaluation of the objective function for a particular Fourier-phase estimate requires only one FFT. Thus employing Fourier-phase values as optimization parameters is certainly competitive with the use of input-pixel values from the standpoint of operations required to compute the gradient. How these two gradient-search formulations compare with respect to convergence properties can only be determined by experimentation. We might expect the Fourier-phase formulation to perform differently since the Fourier-phase parameters are so different in character from and nonlinearly related to the input pixel-value parameters. Use of the Fourier phase for parameters has the added appeal that these are in fact the unknowns in the phase-retrieval problem. As a result the Fourier phase formulation is somewhat more direct and may lend itself to analysis when noise is present.

## 6.6 CONCLUSIONS AND FUTURE WORK

We have shown that the ER algorithm may be interpreted as a constrained steepest-descent algorithm for which the objective function consists of the Fourier-domain error metric as a function of pixel values in the latest estimate. In addition we have proposed three new objective functions for performing phase retrieval using gradient-search methods. These include: 1) use of the summation objective function with

pixel values of the latest estimate as optimization parameters, 2) use of the object-domain error metric with input pixel values as parameters, and 3) use of the object-domain error metric with Fourier-phase values as parameters. Analytic expressions for the gradients for each of these approaches have been derived. The simplicity of these expressions implies that gradient-search methods have the hope of being computationally tractable and even competitive with existing input-output algorithms. The total number of FFTs required to evaluate the objective function and compute the gradient for each of these approaches is shown in Table 6.1.

TABLE 6.1. NUMBER OF FFTs REQUIRED FOR GRADIENT-SEARCH APPROACHES

| Objective Function | #FFTs to evaluate objective function | #FFTs to evaluate gradient |
|---|---|---|
| $e_F^2(g(x))$ | 1 | 2 |
| $e_S^2(g(x))$ | 1 | 2 |
| $e_o^2(g(x))$ | 2 | 5 |
| $e_o^2(\phi(u))$ | 1 | 2 |

Of course extensive experimentation needs to occur to see if the surface contour of each proposed objective function is well suited for gradient-search methods. Surface contour depends upon such things as the intrinsic definition of the objective function, the particular true object, and the amount of noise in the data. The suitability of a particular gradient-search algorithm to a given surface contour manifests itself in the convergence rates of the algorithm. For example a memoryless modified Newton method [6.10] may converge well with the same objective function for which a steepest-descent algorithm stagnates. Software is currently being developed to test for the convergence rates as a function of type of objective function, choice of true object, amount of noise and gradient-search method employed. In

addition, these gradient-search approaches need to be tested in a role that complements current input-output algorithms. Gradient-search approaches could make a significant contribution to the field of phase retrieval, should they consistently provide a mode of escape from any of the various typesof stagnation that have been known to appear with input-output algorithms.

## REFERENCES


6.1 For a list of references see Ref. 4 in D. Kohler and L. Mandel, J. Opt. Soc. Am. 63, 134 (1973).

6.2 R. W. Gerchberg and W. O. Saxton, Optik 35, 237 (1972).

6.3 J. R. Fienup, "Phase Retrieval Algorithms: A Comparison," Applied Optics 21, 2758-2769 (1982).

6.4 J. R. Fienup, "Reconstruction of Objects having Latent Reference Points," J. Opt. Soc. Am. 73, 1421-1426 (1983).

6.5 T. R. Crimmins, "Phase retrieval for discrete functions with support constraints: Summary", OSA Topical meeting on Signal Recovery and Synthesis II, Honolulu, Hawaii (April, 1986).

6.6 R. A. Gonsalves, "Phase retrieval from modulus data," J. Opt. Soc. Am. 66, 961-964 (1976).

6.7 W. O. Saxton, Image Processing in Electron Microscopy (Academic Press, NY 1978).

6.8 R. H. Boucher, "Convergence of Algorithms for Phase Retrieval from Two Intensity Distributions," in International Optical Computing Conference, Proc. SPIE 231, 130-141 (1980).

6.9 J. R. Fienup and C. C. Wackerman, "Phase Retrieval Stagnation Problems and Solutions," Submitted for publication in J. Opt. Soc. Am. A.

6.10 D. G. Luenberger, Linear and Nonlinear Programming (Addison-Wesley, Reading, Massachusetts 1984).

# 7
# MODELING APPROACH TO PHASE RETRIEVAL

The modeling approach is a new method for attempting to solve the phase retrieval problem. In this section we describe the modeling approach in general terms, and then discuss a particular implementation that was attempted.

Let $F(u,v) = |F(u,v)| \exp[i\psi(u,v)]$ be the complex Fourier transform of a particular object. Suppose that either $F(u,v)$ over the entire measurement aperture or $F(u,v)$ over some small area can be modeled by a parameterized function, M:

$$M(u,v,;a,b,...) = |M(u,v;a,b...)| \exp[i\phi(u,v;a,b,...)], \quad (7\text{-}1)$$

where a,b,... are unknown parameters. If we are given only the Fourier modulus, $|F(u,v)|$, then it might be possible to estimate the phase, $\psi(u,v)$, by (1) finding the values of the parameters a,b,... that best fit the modulus of the model, $|M(u,v;a,b,...)|$, to $|F(u,v)|$, and (2) evaluating $\phi(u,v;a,b,...)$ for that set of values of the parameters.

The most difficult part of this approach is finding a model, M, that is suitable.

In a first attempt at using the modeling approach, each small area about the local maxima of the Fourier modulus was modeled using a function taken from the control theory literature. Suppose that contours about a local maximum of the Fourier modulus, at a level 3 dB down from the local maximum, have an elliptical shape, with the major axis of length $w_b$ at an angle $\theta_b$ relative to the u-axis and minor axis $w_d$. Let the local maximum be at location $(u_b,v_b)$ where it has the value

$$A_b = |F(u_b,v_b)| . \quad (7\text{-}2)$$

Also define the distance from a given point $(u,v)$ to the peak $(u_b,v_b)$ as

$$w = [(u-u_b)^2 + (v-v_b)^2]^{1/2} , \tag{7-3}$$

and let

$$\theta = \tan^{-1}[(v-v_b)/(u-u_b)] , \tag{7-4}$$

and

$$w_c = w_b \cos(\theta_b-\theta) + w_d \sin(\theta_b-\theta) . \tag{7-5}$$

Then the model we used for a region about the local maximum is

$$M(w;A_b,\theta_b,w_b,w_d,D) = \frac{A_b w_c^2}{w_c^2 - w^2 + i2ww_cD} \tag{7-6}$$

which has squared modulus

$$|M|^2 = \frac{A_b^2 w_c^4}{(w_c^2 - w^2)^2 + (2ww_cD)^2} \tag{7-7}$$

and phase

$$\phi = -\tan^{-1}[2ww_cD/(w_c^2 - w^2)] . \tag{7-8}$$

Note that the parameters $w_b$, $w_d$ and $\theta_b$ are contained within $w_c$.

These expressions were used in the following way:

(1) A local maximum of the squared Fourier modulus was found.

(2) A curve fit of Eq.(7-7) to the squared Fourier modulus was performed to estimate the unknown parameters.

(3) The phase in that region was computed by Eq.(7-8) using the parameter estimates.

(4) Eq.(7-7) was evaluated using the parameter estimates and subtracted from the squared Fourier modulus, leaving the residual Fourier modulus.

(5) Repeat steps (1) to (4) replacing the squared Fourier modulus with the residual Fourier modulus, until all the major local maxima are accounted for.

(6) Form the net Fourier phase as the sum of all the phase functions obtained in step (3).

(7) Form an image by inverse Fourier transforming the complex function formed from the given Fourier modulus and the net Fourier phase.

Note that for large w, $|M|^2$ approaches zero and $\phi$ in Eq.(7-8) approaches zero, so the model has strong local effect near each local maximum and a weaker effect on neighboring points.

When the procedure was performed for a SAR image of the type used in the digital experiments described in Section 5, the reconstructed image bore no resemblance to the original object. The reason for failure is not totally understood, but we speculate that the model, Eq.(7-6), is not appropriate to the Fourier transforms of SAR images.

If further work along these lines were to be pursued, it would be important to first develop more appropriate models for SAR signal histories.

# 8
# LABORATORY EXPERIMENTS

The objective of Task 3 of the program is to perform laboratory experiments which demonstrate reduced tolerance imaging. These experiments will validate the theoretical developments concerning constraints, measurements, phase retrieval and image reconstruction algorithms, and uniqueness and sensitivity issues under more realistic conditions than is possible in the computer simulations performed under Tasks 1 and 2. The use of real objects, illumination sources, optics, and detectors will place greater demands on the reconstruction algorithms. The quality of the reconstructed images from experimental data will be compared both to images from computer simulations of the experiment and to "ground truth" images collected in the laboratory with a conventional sensor having an equivalent aperture. Experimental parameters (e.g., measurement signal-to-noise ratio, type of shape constraint, sharpness of shape constraint) will be varied for comparision to theoretical and computer simulation results.

Two experiments simulating different types of systems will be performed: an active coherent experiment in the visible and a passive incoherent experiment in the visible or infrared. The results of the active coherent experiment will be useful for predictions concerning SAR and active laser imaging sytems, whereas the passive incoherent experiment will be pertinent to conventional passive and passive interferometric imaging systems. In the active, coherent experiment a target will be illuminated with a laser (with various illumination shapes) and intensity data will be collected in the far-field of the target. Reconstruction algorithms will then be used to determine the phase in the far-field of the target and therefore an image of the target. Thus, this experiment simulates an active, coherent reduced-tolerance sensor which measures intensities only. The optical

and electronic equipment requirements of this experiment have been determined and most of the additional equipment has been purchased, delivered, and tested. An initial setup has been made of part of the equipment in the laboratory. Computer software is under development to collect and process the measurement data. The two experiments will be performed serially, so further planning of the passive experiment is purposely proceeding slowly until the active experiment is approximately one-half complete. Further discussion of both experiments is given in Sections 8.1 and 8.2.

## 8.1 ACTIVE EXPERIMENT

The objective of the active experiment is to demonstrate imaging of a coherently illuminated target from intensity-only measurements made in the far-field. This simulates a sensor having greatly reduced tolerance to the position and quality of its receiving aperture compared to a conventional imaging sensor. The wide range of parameters which must be considered in planning this experiment and which are available to be varied to test theoretical developments and computer simulation results are discussed in Section 8.1.1. Many, but due to finite resources, not all, of the parameters discussed will be exercised in the actual experiments. The experiment design including both optics and electronics is discussed in Section 8.1.2.

In order to greatly increase the range of parameters which could be investigated with a fixed amount of manpower, it was decided at the beginning of the program that an array processor would be purchased which was sufficiently powerful to perform the iterative Fourier transform algorithms needed for image reconstruction and which was compatible with one of ERIM's laboratory computers. (The use of existing VAX facilities would have required time-consuming transfer of large amounts of data and vastly increased computational costs.) A

suitable array processor was ordered, but delays in its delivery have forced the experimental work to fall behind its initial timetable. A less powerful, but similar, array processor has been obtained on loan from the manufacturer (Mercury Computer) until the original order can be shipped. Current work on the laboratory experiments is concerned with writing software to collect and process the measurement data, further setup of optical components, and final equipment purchases.

### 8.1.1. ACTIVE EXPERIMENT PARAMETERS

A wide range of parameters is available to be varied in the active experiment to test theoretical developments and computer simulation results. These parameters must also be carefully controlled to ensure meaningful results. The most important of these parameters are discussed in this section.

The pattern of illumination on the target can be described by its spatial and temporal coherence, shape, sharpness of edges, phase distribution, angle of incidence on the target, and polarization. All of these parameters may affect the quality of the reconstructed image. Equally important, they may take on different values depending on the application being simulated in the laboratory. In an application where the target is actively illuminated by a laser, the spatial (transverse) and temporal (longitudinal) coherence lengths may be less than the target size. The laboratory system can allow illumination with variable coherence lengths either by manipulating the spatial coherence of a gas laser or by using a broader-band dye laser. It is known from ERIM investigations that the illumination pattern shape and sharpness of the edges affects reconstruction algorithm convergence. In applications, the range of illumination shapes and sharpness of edges may be limited by practical considerations on the transmitting aperture. The laboratory system must accept a variety of masks and image them onto the

target through a controllable finite aperture to control illumination shape and sharpness. Since the illumination phase may not be constant over the target in practical applications, the masks will need to be holographic if experimental control of the illumination pattern phase distribution is desired. Applications may be either monostatic or bistatic, so the laboratory system should allow for either. The target will, in most cases, partially depolarize the illumination. The experimental system should be capable of making measurements of the two orthogonal components of the light at the detector.

The target parameters include reflectivity contrast and structure, surface roughness, surface topography (3-D nature of target), motion during the measurement process, and noncoherent background illumination. Practical targets will vary in their roughness, although nearly all will be rough at visible and infrared wavelengths. The experimental system should primarily use rough targets to create real speckle effects. However, it may be useful to use smooth targets (film transparencies in a liquid gate) in setting up the experiment to test and debug the optical and electronic components and software. Real targets are three-dimensional, but to varying degrees. A variety of 3-D objects should be available for the experiment. A practical reduced tolerance sensor may need to cope with target motion during the detector integration time. This effect can be simulated by mounting the target on an motor-driven translation or rotation stage. Any real target will also be noncoherently illuminated from various thermal sources. While this illumination will only add a uniform bias to the far-field measurements, it should be included in the experimental setup.

The propagation path between the target and the sensor can be of such a length as to be either near or far field and can include atmospheric turbulence, scattering (aerosols, fog, smoke), and absorption. In an application, any or all of these effects may be

present. The optical setup of the experiment can allow for insertion or removal of lenses to give either near or far field conditions at the detector. (It must be noted that the size of the speckles in the measurement plane will depend on the sensor distance.) Turbulence with the proper statistical properties is very difficult to simulate in an indoor laboratory. The best approach (and one which allows reproducible turbulence) is to use movable phase plates. These are glass plates with controlled thickness variations. Scattering and absorption are easier to simulate (e.g., with fog chambers, optical narrow band filters). Since their main effects are to reduce signal levels and increase detector bias light levels, they can also be studied as described in the next paragraph on detector parameters.

The most important detector parameters are signal level, type of noise, noise level, background illumination level, spatial and temporal sampling rates, polarization detected, nonlinearities in response, and nonuniformities in response, bias and noise. The values of these parameters are crucial to the viability of any real application. To give useful results, the experimental setup must be able to vary the signal and background illumination levels, simulate various sampling rates, detect orthogonal polarizations, and create nonuniform background illumination levels. The type of noise, noise level, nonuniformities in detector response and noise (pattern noise), and nonlinearities in response will be primarily determined by the detector chosen for the experiment. An important aspect of the experiment design and procedure should be to measure, calibrate, and correct for the effects of these parameters on the measured data. This may be a difficult task and much practical experience will be gained which may be applied to other detectors in the future.

Many applications involving active illumination can be expected to have low signal levels. To adequately simulate these applications, an image intensifier should be used before the detector. Even with a thermal-noise-limited detector, the use of an image intensifier may allow operation in a shot (photon) noise limited mode. The image intensifier will, of course, also have nonuniformities, nonlinearities, and a spatial sampling rate which must be measured and considered in the experimental setup and data analysis.

The effects of speckle are so important to an active coherent experiment that its parameters deserve to be discussed separately. The speckle in the measurement plane will have its size determined chiefly by the size of the illuminated region on the target and by the optics (if any) placed between the target and the sensor. Ideally, the detector must sample the intensity speckle pattern at the Nyquist rate (two samples per speckle) or greater. For a given detector, magnifying optics may be necessary. It may also be desired to investigate the effect of measurements at less than the Nyquist rate. Some speckle reduction techniques include averaging the intensities of images from independent looks (aspect angles) at the target. The experimental setup should be capable of rotating or translating the target in discrete steps to generate these independent looks.

### 8.1.2 ACTIVE EXPERIMENT DESIGN

An optical setup and electronic hardware for performing the active experiment is shown in Fig. 8-1. The laser source is spatially filtered, collimated, and used to illuminate a mask which is imaged onto the target via a beamsplitter. The target is in the front focal plane of lens $L_1$ and the far-field distribution of the light from the target is obtained in the back focal plane. This distribution is imaged with magnification by lenses $L_2$ and $L_3$ (together forming an afocal telescope)

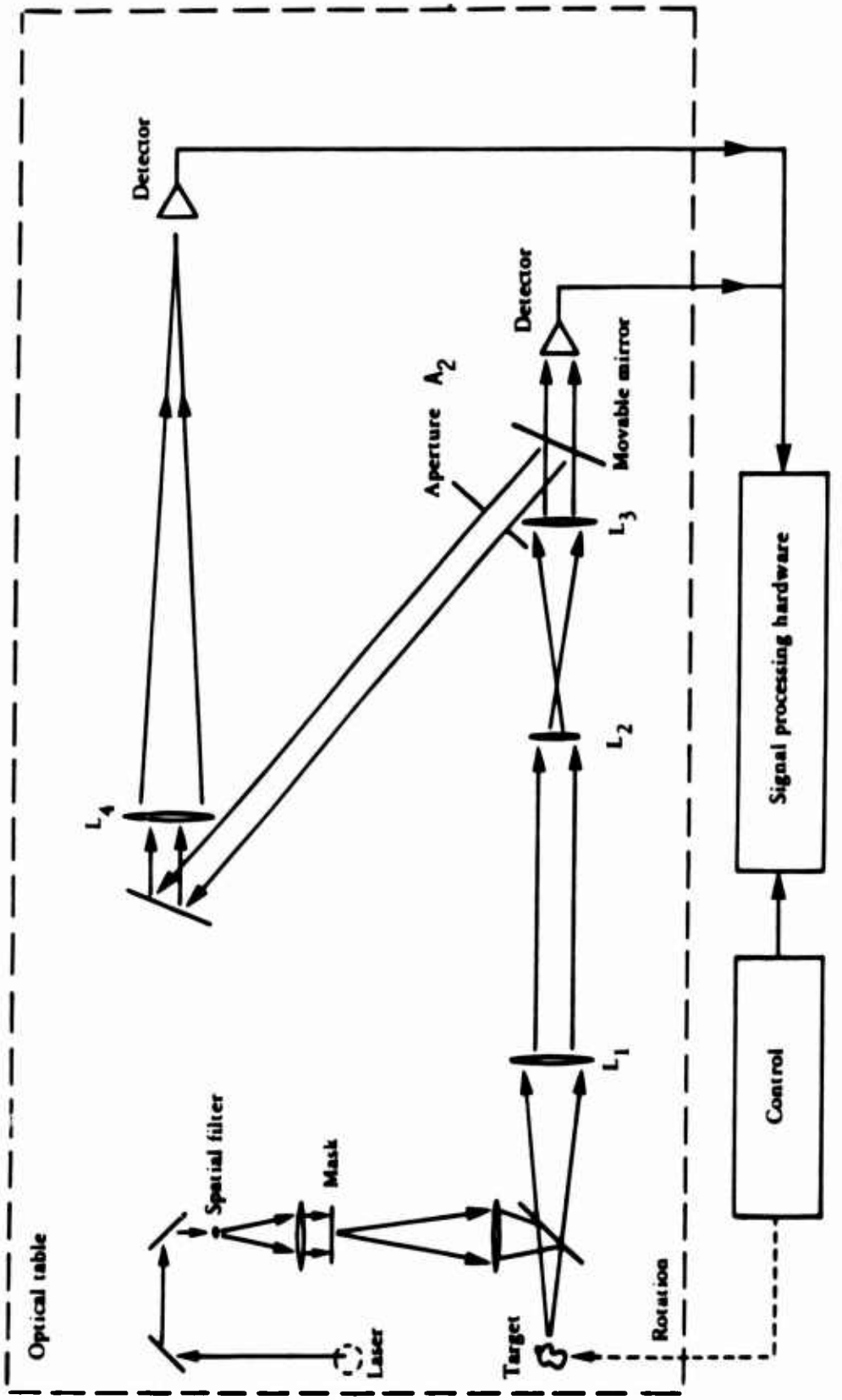



Figure 8-1. (U) Active Experiment Laboratory Setup

onto the detector. Some of this light is redirected by a beamsplitter or a removeable mirror to lens $L_4$ which forms an image of the target on a second detector. Signal processing is used to digitize the detector signals, perform preprocessing of the data, and to perform the phase retrieval and image reconstruction algorithms. A laboratory computer is used to control the experiment and for data storage.

The experiment design outlined above permits most of the parameters discussed in Sec. 8.1.1 to be controlled. For example, it allows the use of various masks and targets, target rotation and translation, noncoherent background illumination of target and detector, variation of detector signal levels, matching of speckle size to detector resolution, detection of orthogonal polarizations, use of an image intensifier, and detection of the target image using a conventional optical sensor. Simple rearrangements of the optical equipment will allow holographic masks to be used, the spatial coherence of the illumination to be varied, monostatic or bistatic illumination, near-field intensities to be measured, and the inclusion of simulated turbulence, scattering, and absorption.

Target, Optics, and Detector Design

For the active (far-field) experiment, the following parameters need to be chosen or determined:

Target diameter, d
Target 1-D space-bandwidth product, N
Lens $L_1$ focal length, $F_1$
Lens $L_1$ aperture diameter, $D_1$
Magnification by lenses $L_2$ and $L_3$, M

Far-field detector diameter, s
Far-field detector element spacing, $\Delta s$
Lens $L_4$ focal length, $F_4$
Target image detector diameter, w
Target image detector element spacing, $\Delta w$
Wavelength, $\lambda$.

These parameters are related by a number of equations. Which parameters can be freely chosen and which are thereby determined depends upon one's point of view. The following explanation is based on the fact that, for experiments performed in this program, most of the parameters will be determined by the capabilities of available detectors.

The allowable target space-bandwidth product (SBP) is obviously limited by the number of detector elements in the far-field detector. Because of the squaring operation inherent in intensity measurements, the spatial frequency spectrum of the detected signal is doubled (relative to amplitude detection) and therefore a detector with 2N elements in each dimension is required to collect the data from which a (possibly complex-valued) target image with a SBP equal to N (in each dimension) can be reconstructed. For current solid state array detectors operating in the visible, the number of detector elements in either dimension varies from about 240 to 490. Therefore, assuming that a square image with N equal to a power of 2 is desired, N will be no greater than 128.

If the target field diameter, d, is chosen, then the resolution required of the optical system to the far-field detector is d/N. The achieved resolution will be $\lambda MF_1/s$. Setting these two expressions equal and using the fact that $s = 2N\Delta s$ gives

$$MF_1 = 2d\Delta s/\lambda. \qquad (8\text{-}1)$$

The detector element spacing, $\Delta s$, therefore has an important effect on several other parameters. Note that the focal length, $F_1$, can be increased to compensate for a larger d and that increases in the magnification factor, M, can be used to decrease $F_1$.

The aperture diameter, $D_1$, of lens $L_1$ must be sufficiently large so that rays leaving the extremes of the target and traveling in directions corresponding to the maximum allowed spatial frequency are not vignetted. The maximum spatial frequency is $s/2\lambda MF_1$ which corresponds to an angle $\theta$ with $\sin\theta = s/2MF_1$. The aperture must be of diameter $d + 2F_1\tan\theta$, so, for $\theta$ small,

$$D_1 = d + s/M. \qquad (8\text{-}2)$$

In the planned experiment, s is less than d and M is greater than 1, so the aperture diameter, $D_1$, is slightly larger than the target diameter, d.

The aperture diameters of lenses $L_2$ and $L_3$ similarly need only be large enough to avoid vignetting and can be easily calculated. The aperture $A_2$ (see Fig. 8-1) in the common focal plane of lenses $L_3$ and $L_4$ must be of diameter equal to the far-field detector diameter, s. This ensures that the image collected by the image detector has the same spatial frequency content as the data collected by the far-field detector.

The magnification from the target to the image detector is $F_4/MF_1$ where $F_4$ is the focal length of lens $L_4$. Since at least 2N samples of the image must be detected, $F_4$ must be at least

$$F_4 = 2MF_1N\Delta w/d \qquad (8\text{-}3)$$

where $\Delta w$ is the detector element spacing of the image detector. By an argument similar to that given above for lens $L_1$, the aperture diameter of lens $L_4$ must be at least s + w.

One possible set of parameters calculated from the above and which represents the current plan for the active experiment is:

Target diameter, d = 25.7 mm
Target space-bandwidth product, N = 128
Lens $L_1$ focal length, $F_1$ = 500 mm
Lens $L_1$ aperture diameter, $D_1$ = 27.0 mm
Magnification by lenses $L_2$ and $L_3$, M = 6
Far-field detector diameter, s = 7.68 mm
Far-field detector element spacing, $\Delta s$ = 30 microns
Lens $L_4$ focal length, $F_4$ = 1000 mm (must be greater than 896 mm)
Target image detector diameter, w = 8.57 mm (for $F_4$ = 1000 mm)
Target image detector element spacing, $\Delta w$ = 30 microns
Wavelength, $\lambda$ = 0.5145 microns.

Light Level Consideration

The optical intensity, I, incident on the far-field detector will be a product of the following factors (estimated or measured values are given where appropriate):

Laser power, 1 Watt
Transmittance of spatial filter, 0.5
Transmittance of mask, 0.5

Reflectivity of target, 0.5

Two-way efficiency of beamspitter, 0.25

Light collection efficiency (assuming diffuse target),
$\pi(s/2MF_1)^2/2\pi = s^2/8M^2F_1^2$

Transmittance of six lenses, $0.9^6$

Gain due to image intensifier (if used), G

Reciprocal of detector area, $1/s^2$

The resulting expression is approximately:

$$I = 2.1\ G/(MF_1)^2 \times 10^{-3}\ \text{Watts}. \qquad (8\text{-}4)$$

Using the values from the previous paragraph gives an intensity, I, of about 2.3 G x $10^{-8}$ Watts/$cm^2$. This can be compared with a laboratory measurement with a system similar to that of Fig. 8-1 (with no intensifier) which gave an intensity of about 3 x $10^{-8}$ Watts/$cm^2$ for a cast metal target. Since current solid state array detectors have noise equivalent intensities of, for example, about 1.4 x $10^{-8}$ Watts/$cm^2$ for the Fairchild CCD3000, either frame integration or an image intensifier with a gain, G, of at least 100 is planned to be used in the active experiment.

## Software Development

Software is currently under development to control the signal processing hardware and to permit digitization of detector signals, preprocessing of the data, and computation of phase retrieval and image reconstruction algorithms. An outline of the functions planned for this experiment control program is given below.

1. General

Program to operate in command file and interactive mode.

Program to save journal of all commands issued and their responses to the user via the terminal including comment lines in order to document work done.

2. Data acquisition

Digitize: Digitize and store image in Imaging Technology, Inc. (ITI) frame buffer with correction for calibrated nonuniformities.

Integrate: Digitize n images and sum in array processor (AP) with/without normalization.

3. Image transfer

Transfer: Move images from ITI to AP and hard disk and between AP and hard disk.

4. Image display

Display: Display AP and hard disk images on ITI.

Notes: (1) Conversion from 32 bit to 8 bit data needed

(2) Many options needed:

Display real, imaginary, magnitude, magnitude-squared, or phase

Apply bias and scale (as in y=ax+b)

Display absolute value

Magnify by 2,4,8,... (specify subimage to be displayed)

Sample to give 256x256 image and display in specified quadrant of ITI display (allows four images to be displayed simultaneously for comparison)

Display any size image in any location of display

(3) Some of these options can be done by altering lookup

tables in ITI

(4) Values above and below the 8 bit range of the ITI should be clipped at 0 and 255

Live/Memeory: Toggle between displaying video incoming to ITI and data in ITI frame buffer.

5. Image algebra (all in AP)

Add: Add two images.

Subtract: Subtract two images.

Multiply: Multiply two images.

Divide: Divide two images with user definable result for divide by zero.

Scale: Add bias and scale image.

Threshold: Hard limit above and below.

Logic operations between binary images.

Magnitude: Find magnitude or magnitude-squared of an image.

Phase: Find phase of an image.

Convert: Change real image to/from complex image.

Print: Print values of specified small part of an image.

Statistics: Find mean, variance of image and magnitude-squared of an image.

Maxmin: Find max and min values of image.

Histogram: Compute histogram of image and display on ITI.

Convolve: Convolve image with a small specified convolution kernel (allows smoothing and other operations on data).

Interpolation: Interpolate from one sample spacing to another.

6. Create images (in AP) for test purposes

Zero: Zero fill an image.

Create: Place a rectangular, circular, or triangular region of

specified complex value at a specified position in an image.

Aperture: Multiply image by a binary rectangular, circular, or triangular aperture located at a specified position.

Noise: Add zero-mean Gaussian noise with specified variance to an image (should specify seed so that same or different set of random numbers can be generated) (Also include uniform and Poisson noise).

7. Image warp

Measure warp: By use of calibrated test patterns, measure image magnification and distortion.

Remap: Resample image to compensate for magnification and distortion.

8. Iterative algorithm

Setup: Allocate and load image domain, Fourier domain, image domain constraint, Fourier magnitude constraint, and buffer arrays in AP.

Iterate: Iterate n times using specified form of iterative algorithm, computing and printing error measures.

Display: Display intermediate results.

Save: Save results.

9. Image error computation

Error measure: Compute normalized root-mean-squared error of complex image or of image magnitude relative to reference object, taking into account intensity scaling and (for complex images) constant phase shift.

10. Help information

Help: Print list of available commands.

Print help information about specified command.

11. Termination

Stop: Complete all commands issued including writing all buffers to hard disk, save journal file, and return to UNIX.

Journal file should also be saved at each step in case of program or system crash.

## 8.2 PASSIVE EXPERIMENT

The objective of the passive experiment is to demonstrate imaging (in the visible or infrared) of a noncoherently illuminated or emitting target from intensity-only or intensity and reduced tolerance phase measurements. Several candidate experiments, currently under consideration, are described below.

### Stellar Speckle Inteferometry

Images of space objects from ground facilities are degraded by the effects of the turbulent atmosphere. One solution to this problem is to use adaptive optics and to correct for the phase distortions of the atmosphere in real time. This solution, of course, requires precise phase measurement and compensation in real time. The reduced tolerance solution is to use (stellar) speckle interferometry which can determine the magnitude of the Fourier transform of the target image despite the turbulent atmosphere. Phase retrieval and image reconstruction algorithms can then be used to form an image of the target. Speckle interferometry has been used by astronomers to image simple objects such as binary stars. The technique has also been used in computer

simulations to image more complicated objects. An experiment demonstrating the use of speckle interferometry to image complicated objects would therefore be a significant step forward in the development of reduced tolerance imaging through turbulence. The experimental setup required could use much of the equipment from the active experiment with the addition of glass plates with small thickness variations to simulate the effect of the turbulent atmosphere.

Passive Synthetic Aperture Imaging

ERIM is currently developing techniques for passive synthetic aperture interferometric imaging of noncoherently illuminated or emitting targets in the visible and infrared. These techniques require accurate alignment and position control of the sensor optics (similar in degree to that required in conventional imaging). The use of reduced tolerance imaging techniques could reduce these accuracy requirements. It would be very appropriate and effective for ERIM to initiate the research to combine these two techniques. The experimental setup required would rely heavily on the equipment used for passive interferomtrc imaging. However, the passive interferometric imaging program plan is such that its experiments will take place in parallel with or after, rather than before, those of the reduced tolerance imaging program, so it may not be feasible to link the two experimental programs together.

Imaging with Phase Diversity

In conventional passive imaging systems where the image is degraded by phase aberrations due either to atmospheric turbulence or to misalignment of segmented optical element arrays, it is known that image quality can be improved by using iterative reconstruction algorithms operating on two 2-D intensity measurements. In the case of turbulence,

these two measurements can be of a best focus image and an intentionally slightly defocused one (that is, a quadratic phase error is intentionally introduced -- phase diversity). For a primary mirror made of segments, the segments may be slightly moved or tilted by a known amount between data collections in the image plane. The use of this approach allows reduced tolerance to atmospheric phase or to accurate positioning and alignment of segmented optics. The equipment required here would again be similar to that used in the active experiment except that piezo-electric actuators controlling multiple segments would be required.

Further planning of the passive experiment is purposely proceeding slowly until the active experiment is approximately one-half complete. If the stellar speckle interferometry experiment is chosen, then the optical equipment requirements are not expected to involve any major purchases beyond those of the active experiment and, in any case, the electronic signal processing hardware already purchased will serve both experiments.

## Appendix A
## PROOF OF THE UNIQUENESS THEOREM

In this appendix the uniqueness theorem presented in Section 3.2.3 is proven.

Let S, T, $v_n$ for $n = 0, \ldots, S-1$, and $q_n$ for $n = 0, \ldots, T-1$ be defined as in Section 3.2.4. Therefore all the $q_n$ referred to in this appendix are reference points. Also, let U, $s_n$ and $w_n$ for $n = 0, \ldots, S-1$ and $k_n$ and $y_n$ for $n = 0, \ldots, T-1$ be defined as in Section 3.2.5. Let $t_n$ be the side of [R(M)] with endpoints $p_n$ and $p_{(n+1) \bmod T}$ (see Figure A-1) and let $u_n = p_{(n+1) \bmod T} - p_n$ for $n = 0, \ldots, T-1$. We note for future reference that for $v, w \in \mathcal{R}^2$, $\langle v, Uv \rangle = 0$, $U^2 v = -v$, $\langle Uv, Uw \rangle = \langle v, w \rangle$, and $\langle v, Uw \rangle = \langle Uv, U^2 w \rangle = -\langle Uv, w \rangle$. The proof of the uniqueness theorem in Section 3.2.3 requires a series of lemmas.

Lemma A-1: R(M) is a mask and R(R(M)) = R(M).

Proof: Suppose it can be shown that every vertex of [R(M)] is opposite some side of [R(M)]. Since at most one vertex can be opposite a given side and the number of vertices equals the number of sides, it would then follow that every side must have a vertex opposite it and therefore no two sides can be parallel, hence R(M) is a mask. Also, since every vertex is opposite a side, R(R(M)) is the set of all vertices of [R(M)] which is equal to the set R(M), hence R(R(M)) = R(M). Thus it suffices to show that every vertex of [R(M)] is opposite some side of [R(M)].

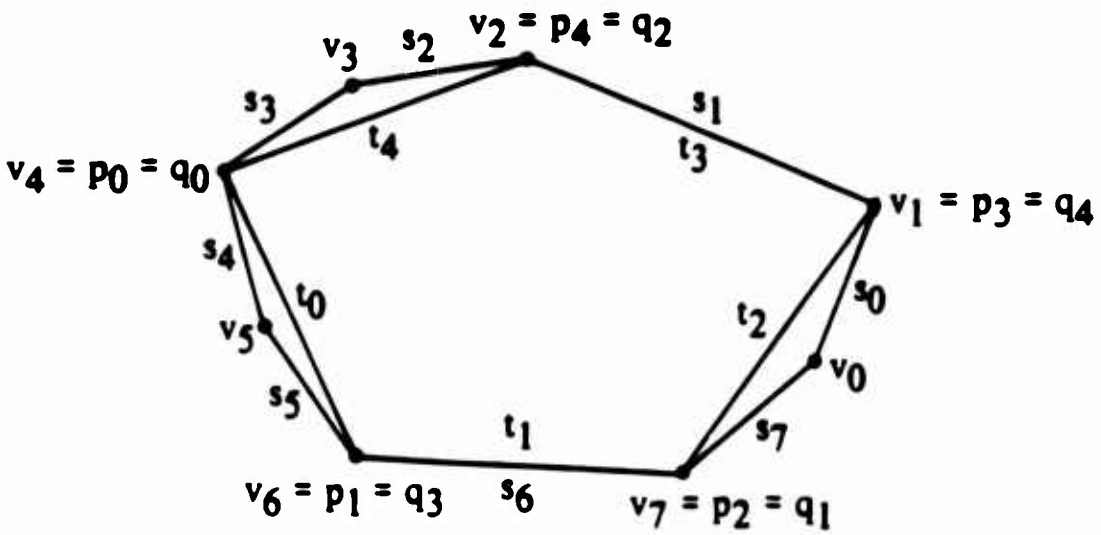



FIGURE A-1. THE SET [R(M)] IS THE CONVEX POLYGON WITH SIDES $t_i$, $i = 0, \ldots, 4$.

The vertices of [R(M)] are $p_m$, $m = 0, \ldots, T-1$. Let m be fixed but arbitrary, $0 \leq m \leq T-1$. Then $p_m$ is also a vertex of [M] and hence $p_m = v_k$ for some k. Let $v_a$ be the vertex of [M] opposite side $s_{(k-1) \bmod S}$ of [M] and let $v_b$ be the vertex of [M] opposite side $s_k$ of [M]. Then $v_a$ and $v_b$ are in R(M) and $v_a = p_n$ for some n. (Refer to Figure A-1 and take $m = 0$. In this case $k = 4$, $a = 7$, $b = 1$ and $n = 2$.) If $v_a$ and $v_b$ are the same vertex, then there can be no side of [M] opposite $v_k$. But $v_k = p_m \in R(M)$ and so $v_k$ must be opposite some side of [M]. Therefore $v_a \neq v_b$. It then follows that $v_b = p_{(n+1) \bmod T}$. It will be shown that $p_m$ is opposite side $t_n$ of [R(M)]. That is, we wish to show that for $0 \leq j \leq T-1$ and $j \neq m$,

$$\langle p_j, Uu_n \rangle < \langle p_m, Uu_n \rangle . \qquad (A-1)$$

Since $v_a$ is opposite side $s_{(k-1) \bmod S}$ of [M] it follows that for $0 \leq i \leq S-1$ and $i \neq a$,

$$\langle Uw_{(k-1) \bmod S}, v_i - v_a \rangle < 0 \qquad (A-2)$$

and since $v_b$ is opposite $s_k$, for $0 \leq i \leq S-1$ and $i \neq b$,

$$\langle Uw_k, v_i - v_b \rangle < 0. \qquad (A-3)$$

Setting $i = b$ in (A-2) we obtain $\langle Uw_{(k-1) \bmod S}, v_b - v_a \rangle < 0$ and thus

$$\langle v_{(k-1)\bmod S} - p_m, Uu_n \rangle = \langle v_{(k-1)\bmod S} - v_k, U(p_{(n+1)\bmod T} - p_n) \rangle$$
$$= -\langle w_{(k-1)\bmod S}, U(v_b - v_a) \rangle$$
$$= \langle Uw_{(k-1)\bmod S}, v_b - v_a \rangle$$
$$< 0. \qquad \text{(A-4)}$$

Setting $i = a$ in (A-3) we obtain $\langle Uw_k, v_a - v_b \rangle < 0$ and thus

$$\langle v_{(k+1)\bmod S} - p_m, Uu_n \rangle = \langle v_{(k+1)\bmod S} - v_k, U(p_{(n+1)\bmod T} - p_n) \rangle$$
$$= \langle w_k, U(v_b - v_a) \rangle$$
$$= \langle Uw_k, v_a - v_b \rangle < 0. \qquad \text{(A-5)}$$

Since $v_{(k-1)\bmod S}$, $v_k (= p_m)$ and $v_{(k+1)\bmod S}$ are distinct vertices of [M], the vectors $v_{(k-1)\bmod S} - p_m$ and $v_{(k+1)\bmod S} - p_m$ are linearly independent. Let $p \in R(M)$, $p \neq v_{(k-1)\bmod S}, p_m, v_{(k+1)\bmod S}$. Then there exist real numbers, $\alpha$ and $\beta$, such that

$$p - p_m = \alpha(v_{(k-1)\bmod S} - p_m) + \beta\,(v_{(k+1)\bmod S} - p_m). \qquad \text{(A-6)}$$

Also, since $\langle v_k, Uw_k \rangle < \langle p, Uw_k \rangle$,

$$0 < \langle p - v_k, Uw_k \rangle$$
$$= \langle p - p_m, U(v_{(k+1)\bmod S} - p_m) \rangle$$
$$= \alpha \langle v_{(k-1)\bmod S} - p_m, U(v_{(k+1)\bmod S} - p_m) \rangle$$
$$+ \beta \langle v_{(k+1)\bmod S} - p_m, U(v_{(k+1)\bmod S} - p_m) \rangle$$
$$= \alpha \langle v_{(k-1)\bmod S} - p_m, U(v_{(k+1)\bmod S} - p_m) \rangle$$

$$= -\alpha \langle w_{(k-1) \bmod S}, U(v_{(k+1) \bmod S} - v_k) \rangle$$

$$= \alpha \langle Uw_{(k-1) \bmod S}, v_{(k+1) \bmod S} - v_k \rangle \ . \qquad \text{(A-7)}$$

Since $\langle Uw_{(k-1) \bmod S}, v_{(k+1) \bmod S} - v_k \rangle > 0$, it follows from (A-7) that $\alpha > 0$. Similarly,

$$0 < \langle p - v_k, Uw_{(k-1) \bmod S} \rangle$$

$$= \langle p - p_m, U(p_m - v_{(k-1) \bmod S}) \rangle$$

$$= \alpha \langle v_{(k-1) \bmod S} - p_m, U(p_m - v_{(k-1) \bmod S}) \rangle$$

$$+ \beta \langle v_{(k+1) \bmod S} - p_m, U(p_m - v_{(k-1) \bmod S}) \rangle$$

$$= -\alpha \langle p_m - v_{(k-1) \bmod S}, U(p_m - v_{(k-1) \bmod S}) \rangle$$

$$+ \beta \langle v_{(k+1) \bmod S} - p_m, U(p_m - v_{(k-1) \bmod S}) \rangle$$

$$= \beta \langle v_{(k+1) \bmod S} - p_m, U(p_m - v_{(k-1) \bmod S}) \rangle$$

$$= \beta \langle v_{(k+1) \bmod S} - v_k, Uw_{(k-1) \bmod S} \rangle \ . \qquad \text{(A-8)}$$

Since $\langle v_{(k+1) \bmod S} - v_k, Uw_{(k-1) \bmod S} \rangle > 0$, it follows from (A-8) that $\beta > 0$. Using (A-4), (A-5), (A-6) and the fact that $\alpha > 0$ and $\beta > 0$, we have, for $p \neq v_{(k-1) \bmod S}, p_m, v_{(k+1) \bmod S}$

$$\langle p - p_m, Uu_n \rangle = \alpha \langle v_{(k-1) \bmod S} - p_m, Uu_n \rangle$$

$$+ \beta \langle v_{(k+1) \bmod S} - p_m, Uu_n \rangle$$

$$< 0. \qquad \text{(A-9)}$$

Inequality (A-1) now follows from (A-4), (A-5) and (A-9). This completes the proof of Lemma A-1.

Lemma A-2: The number, T, of points in R(M) is odd and if $K = (T - 1)/2$ and $0 \leq j \leq T - 1$ then $p_j$, as a vertex of [R(M)], is opposite side $t_{(K+j) \bmod T}$ of [R(M)].

Proof: It follows from Lemma A-1 that every side of [R(M)] has exactly one vertex of [R(M)] opposite from it and every vertex, i.e., every point in R(M), is opposite from exactly one side. Thus there is a positive integer $K \leq T - 2$ such that $p_{K+1}$ is opposite side $t_0$. Then $p_1$ is opposite side $t_{K+1}$ and $p_2$ is opposite side $t_{(K+2) \bmod T}$. More generally, $p_j$ is opposite side $t_{(K+j) \bmod T}$. Setting $j = T - K$ we obtain $p_{T-K}$ is opposite $t_0$. But $p_{K+1}$ is opposite $t_0$. Therefore $T - K = K + 1$ or $T = 2K + 1$. This completes the proof of Lemma A-2.

Lemma A-3: For $x \in [M]$, $x \neq q_j, q_{(j+1) \bmod T}$, $j = 0, \ldots, T - 1$, $\langle q_j, y_j \rangle < \langle x, y_j \rangle < \langle q_{(j+1) \bmod T}, y_j \rangle$.

Proof: It suffices to show that the inequalities hold for all vertices v of [M], $v \neq q_j, q_{(j+1) \bmod T}$. Let j be fixed but arbitrary. For convenience let $m = (k_j - 1) \bmod S$ and $n = k_{(j+1) \bmod T}$. Then $y_j = Uw_m - Uw_n$.

First we will show that $q_{(j+1) \bmod T}$, as a vertex of [M], is opposite side $s_m$ of [M]. Let $v_a$ be the vertex of [M] opposite side $s_m$ of [M] and let $v_b$ be opposite $s_{(m+1) \bmod S}$. Then $v_a$ and $v_b$ are in

R(M) and $v_a = p_k$ for some k. (Refer to Figure A-1 and take j = 4. Then $k_j = 1$, m = 0, (j + 1)mod T = 0, $n = k_0 = 4$, a = 4, b = 6 and k = 0.) By the argument in the proof of Lemma A-1, $v_b = p_{(k+1)\text{mod } T}$ and $p_{(jK)\text{mod } T}$ ( $= q_j = v_{(m+1)\text{mod } S}$), as a vertex of [R(M)], is opposite side $t_k$ of [R(M)]. By Lemma A-2, $p_{(jK)\text{mod } T}$ is opposite $t_{(j+1)K \text{ mod } T}$. Hence, by Lemma A-1, $t_k = t_{(j+1)K \text{ mod } T}$ and $k = (j+1)K \text{ mod } T$. Thus $q_{(j+1)\text{mod } T} = p_{(j+1)K \text{ mod } T} = p_k = v_a$, and therefore $q_{(j+1)\text{mod } T}$, as a vertex of [M], is opposite side $s_m$ of [M].

By a similar argument it can be shown that $q_j$, as a vertex of [M], is opposite side $s_n$ of [M].

Since $q_{(j+1)\text{mod } T}$ is opposite $s_m$ and $v \notin q_{(j+1)\text{mod } T}$, it follows that $\langle q_{(j+1)\text{mod } T}, Uw_m \rangle > \langle v, Uw_m \rangle$ or $\langle q_{(j+1)\text{mod } T} - v, Uw_m \rangle > 0$. Also, since $v \in [M]$, $\langle v - v_n, Uw_n \rangle \geq 0$. Therefore,

$$\begin{aligned}
\langle q_{(j+1)\text{mod } T}, y_j \rangle - \langle v, y_j \rangle &= \langle q_{(j+1)\text{mod } T} - v, y_j \rangle \\
&= \langle q_{(j+1)\text{mod } T} - v, Uw_m - Uw_n \rangle \\
&= \langle q_{(j+1)\text{mod } T} - v, Uw_m \rangle - \langle q_{(j+1)\text{mod } T} - v, Uw_n \rangle \\
&= \langle q_{(j+1)\text{mod } T} - v, Uw_m \rangle - \langle v_n - v, Uw_n \rangle \\
&= \langle q_{(j+1)\text{mod } T} - v, Uw_m \rangle + \langle v - v_n, Uw_n \rangle \\
&> 0.
\end{aligned} \tag{A-10}$$

Since $q_j$ is opposite $s_n$ and $v \neq q_j$, $\langle q_j, Uw_n \rangle > \langle v, Uw_n \rangle$ or $\langle q_j - v, Uw_n \rangle > 0$. Also, since $v \in [M]$, $\langle v - v_{(m+1) \bmod S}, Uw_m \rangle \geq 0$. Therefore

$$\begin{aligned}
\langle v, y_j \rangle - \langle q_j, y_j \rangle &= \langle v - q_j, y_j \rangle \\
&= \langle v - q_j, Uw_m - Uw_n \rangle \\
&= \langle v - q_j, Uw_m \rangle + \langle q_j - v, Uw_n \rangle \\
&= \langle v - v_{k_j}, Uw_m \rangle + \langle q_j - v, Uw_n \rangle \\
&= \langle v - v_{(m+1) \bmod S}, Uw_m \rangle + \langle q_j - v, Uw_n \rangle \\
&> 0. \qquad \text{(A-11)}
\end{aligned}$$

It now follows from (A-10) and (A-11) that $\langle q_j, y_j \rangle < \langle v, y_j \rangle < \langle q_{(j+1) \bmod T}, y_j \rangle$. This completes the proof of Lemma A-3.

In the remainder of this appendix, all modulo arithmetic will be mod T. For convenience, we define $m \oplus n = (m + n) \bmod T$.

The next lemma asserts that $q_j$ and $q_{j \oplus 1}$ have unique separation in M.

Lemma A-4: For $0 \leq j \leq T - 1$, if $x_1, x_2 \in M$ and $x_1 - x_2 = q_{j \oplus 1} - q_j$, then $x_1 = q_{j \oplus 1}$ and $x_2 = q_j$.

Proof: If either $x_1 \neq q_{j \oplus 1}$ or $x_2 \neq q_j$ then it follows from Lemma A-3 that

$$\langle x_1 - x_2, y_j \rangle = \langle x_1, y_j \rangle - \langle x_2, y_j \rangle$$
$$< \langle q_j \oplus 1, y_j \rangle - \langle q_j, y_j \rangle$$
$$= \langle q_j \oplus 1 - q_j, y_j \rangle \qquad \text{(A-12)}$$

which contradicts the assumption that $x_1 - x_2 = q_j \oplus 1 - q_j$. Therefore $x_1 = q_j \oplus 1$ and $x_2 = q_j$. This completes the proof of Lemma A-4.

Let g and h be complex-valued functions on $Z^2$ and let $g * h$ denote the convolution of g and h. That is, if $f = g * h$ then

$$f(x) = \sum_{u \in Z^2} g(u)h(x - u). \qquad \text{(A-13)}$$

We define $\mathcal{S}(g) + \mathcal{S}(h) = \{ x + y: x \in \mathcal{S}(g) \text{ and } y \in \mathcal{S}(h) \}$. The following Lemma is fundamental.

Lemma A-5: If $f = g * h$ then $[\mathcal{S}(f)] = [\mathcal{S}(g) + \mathcal{S}(h)]$.

Proof: It follows from (A-13) that $\mathcal{S}(f) \subseteq \mathcal{S}(g) + \mathcal{S}(h)$ hence $[\mathcal{S}(f)] \subseteq [\mathcal{S}(g) + \mathcal{S}(h)]$. It remains to show that $[\mathcal{S}(g) + \mathcal{S}(h)] \subseteq [\mathcal{S}(f)]$.

Let x be a vertex of $[\mathcal{S}(g) + \mathcal{S}(h)]$. Then there exists a $y \in \mathcal{R}^2$ such that for $x' \in \mathcal{S}(g) + \mathcal{S}(h)$ and $x' \neq x$,

$$\langle x', y \rangle < \langle x, y \rangle. \qquad \text{(A-14)}$$

Also, since x is a vertex of $[\mathcal{S}(g) + \mathcal{S}(h)]$, it follows that $x \in \mathcal{S}(g) + \mathcal{S}(h)$ and hence there exist $x_1 \in \mathcal{S}(g)$ and $x_2 \in \mathcal{S}(h)$

such that $x = x_1 + x_2$. We will show that this decomposition of x is unique. Suppose $x = x'_1 + x'_2$ with $x'_1 \in S(g)$ and $x'_2 \in S(h)$. Then

$$\begin{aligned} \langle x_1, y \rangle + \langle x_2, y \rangle &= \langle x_1 + x_2, y \rangle \\ &= \langle x, y \rangle \\ &= \langle x'_1 + x'_2, y \rangle \\ &= \langle x'_1, y \rangle + \langle x'_2, y \rangle \quad . \end{aligned} \tag{A-15}$$

Therefore, either $\langle x'_1, y \rangle \geq \langle x_1, y \rangle$ or $\langle x'_2, y \rangle \geq \langle x_2, y \rangle$ or both. Suppose $\langle x'_1, y \rangle \geq \langle x_1, y \rangle$. Let $x' = x'_1 + x_2$. Then $x' \in S(g) + S(h)$ and

$$\begin{aligned} \langle x', y \rangle &= \langle x'_1, y \rangle + \langle x_2, y \rangle \\ &\geq \langle x_1, y \rangle + \langle x_2, y \rangle \\ &= \langle x, y \rangle \quad . \end{aligned} \tag{A-16}$$

therefore, by inequality (A-14), $x' = x$ which implies that $x'_1 = x_1$ and hence $x'_2 = x_2$. If $\langle x'_2, y \rangle \geq \langle x_2, y \rangle$ a similar argument leads to the same conclusion. Therefore the decomposition $x = x_1 + x_2$ with $x_1 \in S(g)$ and $x_2 \in S(h)$ is unique. Now suppose for a particular $u_0 \in Z^2$, $g(u_0)\, h(x - u_0) \neq 0$. Then $u_0 \in S(g)$, $x - u_0 \in S(h)$ and $x = u_0 + (x - u_0)$. By the uniqueness of the decomposition of x it follows that $u_0 = x_1$ and $x - u_0 = x_2$. Therefore, $f(x) = g(x_1)\, h(x_2) \neq 0$ and $x \in S(f)$. Since x was an arbitrary vertex of $[S(g) + S(h)]$, it follows that all the vertices of $[S(g) + S(h)]$, are in $S(f)$ and therefore $[S(g) + S(h)] \subseteq [S(f)]$. This completes the proof of Lemma A-5.

We are now ready to prove the theorem.

Proof of Theorem: Since $r = r_1$, it follows from the results of Bruck and Sodin [3.3] that there exist functions g and h with finite supports and a vector $d \in Z^2$ such that $f = g * h$ and $f_1(x) = g * h_1(x - d)$ where $h_1(x) = h^*(-x)$ for $x \in Z^2$.

We have $R(M) \subseteq \mathcal{S}(f) \subseteq \mathcal{S}(g) + \mathcal{S}(h)$. Therefore there exist $a_0, \ldots, a_{T-1} \in \mathcal{S}(g)$ and $b_0, \ldots, b_{T-1} \in \mathcal{S}(h)$ such that

$$q_j = a_j + b_j, \; j = 0, \ldots, T - 1 \; . \tag{A-17}$$

Now let j be fixed but arbitrary and let $x \in \mathcal{S}(g)$, $x \neq a_j$. We will show that $\langle a_j, y_j \rangle < \langle x, y_j \rangle$. Suppose to the contrary that $\langle x, y_j \rangle \leq \langle a_j, y_j \rangle$. Let $x' = x + b_j$. Then, using Lemma A-5, $x' \in \mathcal{S}(g) + \mathcal{S}(h) \subseteq [\mathcal{S}(g) + \mathcal{S}(h)] = [\mathcal{S}(f)] \subseteq [M]$. Also,

$$\begin{aligned} \langle x', y_j \rangle &= \langle x, y_j \rangle + \langle b_j, y_j \rangle \\ &\leq \langle a_j, y_j \rangle + \langle b_j, y_j \rangle \\ &= \langle q_j, y_j \rangle \; . \end{aligned} \tag{A-18}$$

It now follows from Lemma A-3 that $x' = q_j$ which implies that $x = a_j$ contradicting the assumption that $x \neq a_j$. Therefore $\langle a_j, y_j \rangle < \langle x, y_j \rangle$. By a similar argument it can be shown that if $x \in \mathcal{S}(g)$ and $x \neq a_j \oplus 1$, then $\langle x, y_j \rangle < \langle a_j \oplus 1, y_j \rangle$. Thus, if $x \in \mathcal{S}(g)$ and $x \neq a_j, a_j \oplus 1$, then

$$\langle a_j, y_j \rangle < \langle x, y_j \rangle < \langle a_j \oplus 1, y_j \rangle \; . \tag{A-19}$$

Also, by similar arguments, it can be shown that if $x \in \mathcal{S}(h)$ and $x \neq b_j, b_j \oplus 1$ then

$$\langle b_j, y_j \rangle < \langle x, y_j \rangle < \langle b_{j \oplus 1}, y_j \rangle . \qquad \text{(A-20)}$$

Let $S(g) - S(h) = \{x - y: \ x \in S(g) \text{ and } y \in S(h)\}$. Since $S(h_1) = -S(h)$, $S(g) + S(h_1) = S(g) - S(h)$. Also, since $f_1(x) = g * h_1(x - d)$, it follows from Lemma A-5 that

$$[S(f_1)] = [S(g) - S(h)] + d. \qquad \text{(A-21)}$$

We will show that

$$a_j - b_{j \oplus 1} + d \in M. \qquad \text{(A-22)}$$

We have $a_j - b_{j \oplus 1} \in S(g) - S(h)$. Let j be fixed but arbitrary and let $x \in S(g) - S(h)$, $x \neq a_j - b_{j \oplus 1}$. Then there exist $x_1 \in S(g)$ and $x_2 \in S(h)$ such that $x = x_1 - x_2$. Since $x \neq a_j - b_{j \oplus 1}$, either $x_1 \neq a_j$ or $x_2 \neq b_{j \oplus 1}$ or both. In any case it follows from (A-19) and (A-20) that

$$\begin{aligned} \langle x, y_j \rangle &= \langle x_1, y_j \rangle - \langle x_2, y_j \rangle \\ &> \langle a_j, y_j \rangle - \langle b_{j \oplus 1}, y_j \rangle \\ &= \langle a_j - b_{j \oplus 1}, y_j \rangle \quad . \end{aligned} \qquad \text{(A-23)}$$

Therefore, $a_j - b_{j \oplus 1}$ is a vertex of $[S(g) - S(h)]$ and by (A-21), $a_j - b_{j \oplus 1} + d$ is a vertex of $[S(f_1)]$. Therefore $a_j - b_{j \oplus 1} + d \in S(f_1) \subseteq M$ and (A-22) follows.

By a similar argument it can be shown that

$$a_{j \oplus 1} - b_j + d \in M. \qquad \text{(A-24)}$$

Now

$$(a_j - b_{j \oplus 1} + d) - (a_{j \oplus 1} - b_j + d) = (a_j + b_j) - (a_{j \oplus 1} + b_{j \oplus 1})$$

$$= q_j - q_{j \oplus 1}. \qquad \text{(A-25)}$$

By Lemma A-4, $a_j - b_{j \oplus 1} + d = q_j = a_j + b_j$ from which we obtain

$$b_j + b_{j \oplus 1} = d. \qquad \text{(A-26)}$$

From $b_j + b_{j \oplus 1} = d$ and $b_{j \oplus 1} + b_{j \oplus 2} = d$ we obtain $b_j = b_{j \oplus 2}$. Since by Lemma A-2 T is odd, it now follows that $b_0 = b_1 = \ldots = b_{T-1}$ and, using (A-26), we obtain

$$b_0 = b_1 = \ldots = b_{T-1} = d/2. \qquad \text{(A-27)}$$

From (A-20) and (A-27) we obtain $\mathcal{S}(h) = \{d/2\}$. Therefore, for $x \in Z^2$,

$$f(x) = g * h\,(x)$$

$$= h(d/2)g(x - d/2). \qquad \text{(A-28)}$$

If $h(d/2) = 0$ then f would be identically zero, contradicting the assumption that $R(M) \subseteq \mathcal{S}(f)$. Therefore, $h(d/2) \neq 0$. Now, for $x \in Z^2$,

$$f_1(x) = g * h_1(x - d)$$

$$= h^*(d/2)g(x - d + d/2)$$

$$= h^*(d/2)g(x - d/2)$$

$$= \alpha f(x), \qquad \text{(A-29)}$$

where

$$\alpha = \frac{h^*(d/2)}{h(d/2)}.$$

Since $|\alpha| = 1$, this completes the proof of the theorem.

## Appendix B

## PROOF OF PROGRAM FOR IMPLEMENTING RECONSTRUCTION ALGORITHM

It will be shown in this appendix that the program presented at the end of Section 3.2.4 computes $f(q_n)$ for $n = T, \ldots, N-1$. It will be assumed that $f(q_n)$ for $n = 0, \ldots, T-1$ has been computed as described in Section 3.2.4. Then, since $0 \leq m_n \leq T-1$, $f(m_n)$ has been computed for $n = T, \ldots, N-1$.

For $T \leq n \leq N-1$ we have

$$r(q_n - q_{m_n}) = \sum_{y \in Z^2} f(y) f^*(y - q_n + q_{m_n}). \tag{B-1}$$

If $y \in S(f)$ and $y - q_n + q_{m_n} \in S(f)$ then, since $S(f) \subseteq M$, it follows that $y \in M$ and $y - q_n + q_{m_n} \in M$ or, equivalently, $y \in M + q_n - q_{m_n}$. Therefore,

$$y \in M \cap (M + q_n - q_{m_n}) \subseteq \{ q_0, \ldots, q_n \}. \tag{B-2}$$

Hence,

$$r(q_n - q_{m_n}) = \sum_{k=0}^{n} f(q_k) f^*(q_k - q_n + q_{m_n})$$

$$= f(q_n) f^*(q_{m_n}) + \sum_{k=0}^{n-1} f(q_k) f^*(q_k - q_n + q_{m_n}). \tag{B-3}$$

Thus,

$$f(q_n) = \frac{1}{f^*(q_{m_n})}\left[r(q_n - q_{m_n}) - \sum_{k=0}^{n-1} f(q_k)f^*(q_k - q_n + q_{m_n})\right]. \quad (B\text{-}4)$$

An induction argument will be used to prove that $f(q_n)$ is computed correctly for $n = T, \ldots, N-1$. The induction hypothesis is

$$H(n): \quad f(q_j) \text{ is computed correctly for } 0 \leq j \leq n. \quad (B\text{-}5)$$

By the derivation in Section 3.2.4, $H(T-1)$ is true. Now assume $T \leq n \leq N-1$ and $H(n-1)$ is true. We want to show that $H(n)$ is true. It suffices to show that $f(q_n)$ is computed correctly. Let all variables have the values that they have at Step 2 of the pass through the loop in which $f(q_n)$ is computed. It must be shown that all values of f appearing in the right-hand side of Eq. (B-4) are correct. Since, by assumption, $H(n-1)$ is true it follows that $f(q_k)$, $k = 0, \ldots, n-1$, have the correct values. Also, as mentioned above, $f(q_{m_n})$ has the correct value. Now let $x = q_k - q_n + q_{m_n}$ with $0 \leq k \leq n-1$. If $x \notin M$, then $f(x) = 0$. In this case, since $\mathcal{S}(f) \subseteq M$, it follows that $x \notin \mathcal{S}(f)$ and therefore 0 is the correct value for $f(x)$. Now assume $x \in M$. We have $x + q_n - q_{m_n} = q_k \in M$ and therefore $x \in M - q_n + q_{m_n}$. Hence $x \in M \cap (M - q_n + q_{m_n}) \subseteq \{q_0, \ldots, q_{n-1}\}$. It now follows from the induction hypothesis, $H(n-1)$, that $f(x)$ has the correct value when the value for $f(q_n)$ is computed. Therefore $f(q_n)$ is computed correctly and $H(n)$ is true. By induction it follows that $H(N-1)$ is true and hence $f(n)$ is computed correctly for $n = 0, \ldots, N-1$.

## Appendix C

## PROOF OF THE ALGORITHM FOR GENERATING RECONSTRUCTION ALGORITHMS

In this appendix it will be shown that the program in Section 3.2.5 generates a reconstruction algorithm. First, it will be shown that the loop is not infinite and hence the program produces sequences $q_T, \ldots, q_{N-1}$ and $m_T, \ldots, m_{N-1}$. Secondly, it will be shown that if $q = (q_0, \ldots, q_{N-1})$ and $m = (m_T, \ldots, m_{N-1})$ then $(q, m)$ is a reconstruction algorithm.

In the following, 0 will be used to denote both the number zero and the origin of $\mathcal{R}^2$. Context should prevent any confusion.

If $x, y, z \in \mathcal{R}^2$, let $[x, y, z]$ denote their convex hull in $\mathcal{R}^2$. If $x$, $y$ and $z$ are non-collinear then the interior of $[x, y, z]$ is given by

$$\text{int}[x, y, z] = \{ax + by + cz: \ a, b, c > 0 \text{ and } a + b + c = 1\}. \tag{C-1}$$

Then $0 \in \text{int}[x, y, z]$ if and only if $x$, $y$ and $z$ are non-collinear and there exist strictly positive numbers $a$, $b$, $c$ such that $ax + by + cz = 0$. (The sum of $a$, $b$ and $c$ can be made equal to 1 by dividing each of these numbers by their sum if necessary.)

The following lemma will be needed.

Lemma C-1: If $u_n, v_n \in \mathcal{R}^2$, $n = 1, 2, 3$, $\langle u_n, v_n \rangle > 0$ and $\langle u_n, v_m \rangle < 0$ for $n \neq m$, then $0 \in \text{int}[u_1, u_2, u_3]$ and $0 \in \text{int}[v_1, v_2, v_3]$.

Proof: By symmetry it suffices to show that $0 \in \text{int}[\mu_1, \mu_2, \mu_3]$.

First, we will show that $\mu_1$, $\mu_2$ and $\mu_3$ are non-collinear. If they were collinear then one of them would be in the convex hull of the other two. Say $\mu_1$ is in the convex hull of $\mu_2$ and $\mu_3$. Then there would exist numbers $\tau_2$ and $\tau_3$ such that $\tau_2, \tau_3 \geq 0$, $\tau_2 + \tau_3 = 1$ and $\mu_1 = \tau_2\mu_2 + \tau_3\mu_3$. But then $\langle \mu_1, \nu_1 \rangle = \langle \tau_2\mu_2 + \tau_3\mu_3, \nu_1 \rangle = \tau_2 \langle \mu_2, \nu_1 \rangle + \tau_3 \langle \mu_3, \nu_1 \rangle < 0$, contradicting the assumption that $\langle \mu_1, \nu_1 \rangle > 0$. Therefore $\mu_1$, $\mu_2$ and $\mu_3$ are non-collinear.

Since any three vectors in $\mathcal{R}^2$ are linearly dependent, there exist three numbers $\sigma_1$, $\sigma_2$ and $\sigma_3$, not all zero, such that

$$\sigma_1\mu_1 + \sigma_2\mu_2 + \sigma_3\mu_3 = 0. \tag{C-2}$$

Since $\langle \mu_n, \nu_n \rangle > 0$, $\mu_n \neq 0$, $n = 1, 2, 3$. Therefore no two of the $\sigma_n$ can be zero.

Now suppose $\sigma_1 = 0$. Then $\sigma_2 \neq 0 \neq \sigma_3$ and it follows from (C-2) that

$$\mu_2 = -\frac{\sigma_3}{\sigma_2}\mu_3, \tag{C-3}$$

hence

$$0 < \langle \mu_2, \nu_2 \rangle = -\frac{\sigma_3}{\sigma_2} \langle \mu_3, \nu_2 \rangle. \tag{C-4}$$

Since $\langle \mu_3, \nu_2 \rangle < 0$, it follows from (C-4) that

$$-\frac{\sigma_3}{\sigma_2} < 0. \tag{C-5}$$

Also,

$$0 > \langle \mu_2, \nu_1 \rangle = -\frac{\sigma_3}{\sigma_2} \langle \mu_3, \nu_1 \rangle , \tag{C-6}$$

and since $\langle \mu_3, \nu_1 \rangle < 0$, it follows from (C-6) that $-\sigma_3/\sigma_2 > 0$ which contradicts (C-5). Therefore $\sigma_1 \neq 0$ and by symmetry, $\sigma_2 \neq 0 \neq \sigma_3$. By multiplying the $\sigma_n$'s by -1 if necessary, we may assume that $\sigma_1 > 0$. Now

$$\begin{aligned} &\sigma_1 \langle \mu_1, \nu_1 \rangle + \sigma_2 \langle \mu_2, \nu_1 \rangle + \sigma_3 \langle \mu_3, \nu_1 \rangle \\ &\quad = \langle \sigma_1\mu_1 + \sigma_2\mu_2 + \sigma_3\mu_3, \nu_1 \rangle \\ &\quad = 0. \end{aligned} \tag{C-7}$$

Since $\sigma_1 > 0$ and $\langle \mu_1, \nu_1 \rangle > 0$,

$$\sigma_2 \langle \mu_2, \nu_1 \rangle + \sigma_3 \langle \mu_3, \nu_1 \rangle = -\sigma_1 \langle \mu_1, \nu_1 \rangle < 0. \tag{C-8}$$

Since $\langle \mu_2, \nu_1 \rangle < 0$ and $\langle \mu_3, \nu_1 \rangle < 0$, at least one of the numbers $\sigma_2$ and $\sigma_3$ must be strictly positive. By symmetry, we may assume without loss of generality that $\sigma_2 > 0$. Now

$$\begin{aligned} &\sigma_1 \langle \mu_1, \nu_3 \rangle + \sigma_2 \langle \mu_2, \nu_3 \rangle + \sigma_3 \langle \mu_3, \nu_3 \rangle \\ &\quad = \langle \sigma_1\mu_1 + \sigma_2\mu_2 + \sigma_3\mu_3, \nu_3 \rangle \\ &\quad = 0. \end{aligned} \tag{C-9}$$

Since $\sigma_1 > 0$, $\sigma_2 > 0$, $\langle \mu_1, \nu_3 \rangle < 0$ and $\langle \mu_2, \nu_3 \rangle < 0$, it follows that

$$\sigma_3 \langle \mu_3, \nu_3 \rangle = -\sigma_1 \langle \mu_1, \nu_3 \rangle - \sigma_2 \langle \mu_2, \nu_3 \rangle$$

$$> 0. \qquad \text{(C-10)}$$

Since $\langle \mu_3, \nu_3 \rangle > 0$, it follows that $\sigma_3 > 0$. By the comment preceeding the lemma, it now follows that $0 \in \text{int}[\mu_1, \mu_2, \mu_3]$. This completes the proof of Lemma C-1.

One more lemma is needed before proving that the loop is not infinite. As in Appendix A, we define $j \oplus k = (j + k) \bmod T$.

Lemma C-2: For $j = 0, \ldots, T - 1$ and $k = 2, \ldots, T - 1$, $0 \in \text{int}[y_j, y_{j \oplus 1}, (-1)^k y_{j \oplus k}]$.

Proof: The proof will be by induction on k. Let $k = 2$. We want to show that $0 \in \text{int}[y_j, y_{j \oplus 1}, y_{j \oplus 2}]$. Let

$$\mu_1 = q_{j \oplus 1} - q_{j \oplus 3}, \ \mu_2 = q_{j \oplus 2} - q_{j \oplus 1}, \ \mu_3 = q_j - q_{j \oplus 2},$$

$$\nu_1 = y_j, \ \nu_2 = y_{j \oplus 1}, \ \nu_3 = y_{j \oplus 2} \ . \qquad \text{(C-11)}$$

By Lemma A-3, $\langle \mu_n, \nu_n \rangle > 0$ and $\langle \mu_n, \nu_m \rangle < 0$ for $n \neq m$. Therefore, by Lemma C-1, $0 \in \text{int}[\nu_1, \nu_2, \nu_3] = \text{int}[y_j, y_{j \oplus 1}, y_{j \oplus 2}]$.

Now let $3 \leq k \leq T - 1$ and assume the lemma is true for $k - 1$. We want to show that $0 \in \text{int}[y_j, y_{j \oplus 1}, (-1)^k y_{j \oplus k}]$. We have shown that $0 \in \text{int}[y_j, y_{j \oplus 1}, y_{j \oplus 2}]$ and therefore there exist strictly positive numbers $\sigma_1, \sigma_2 \sigma_3$ such that

$$\sigma_1 y_j + \sigma_2 y_{j \oplus 1} + \sigma_3 y_{j \oplus 2} = 0. \qquad \text{(C-12)}$$

Applying the lemma for k - 1 with j replaced by $j \oplus 1$, we have $0 \in \text{int}[y_{j \oplus 1}, y_{j \oplus 2}, (-1)^{k-1} y_{j \oplus k}]$ and therefore there exist strictly positive numbers, $\tau_1, \tau_2, \tau_3$ such that

$$\tau_1 y_{j \oplus 1} + \tau_2 y_{j \oplus 2} + \tau_3 (-1)^{k-1} y_{j \oplus k} = 0. \qquad \text{(C-13)}$$

Multiplying (C-13) by $\sigma_3/\tau_2$ and subtracting the result from (C-12) we obtain

$$\lambda_1 y_j + \lambda_2 y_{j \oplus 1} + \lambda_3 (-1)^k y_{j \oplus k} = 0, \qquad \text{(C-14)}$$

where $\lambda_1 = \sigma_1$, $\lambda_2 = \sigma_2 - \sigma_3 \tau_1/\tau_2$ and $\lambda_3 = \sigma_3 \tau_3/\tau_2$. We have $\lambda_1 > 0$ and $\lambda_3 > 0$. We will show that $\lambda_2 > 0$. From (C-14) we obtain

$$\lambda_2 y_{j \oplus 1} = -\lambda_1 y_j + \lambda_3 (-1)^{k-1} y_{j \oplus k}. \qquad \text{(C-15)}$$

We consider two cases.

Case 1: k is odd. Then k - 1 is even and by (C-15)

$$\lambda_2 y_{j \oplus 1} = -\lambda_1 y_j + \lambda_3 y_{j \oplus k}, \qquad \text{(C-16)}$$

and, using Lemma A-3,

$$\begin{aligned} &\lambda_2 \langle q_{j \oplus k \oplus 1} - q_{j \oplus 1}, y_{j \oplus 1} \rangle \\ &= -\lambda_1 \langle q_{j \oplus k \oplus 1} - q_{j \oplus 1}, y_j \rangle + \lambda_3 \langle q_{j \oplus k \oplus 1} - q_{j \oplus 1}, y_{j \oplus k} \rangle \\ &= \lambda_1 \langle q_{j \oplus 1} - q_{j \oplus k \oplus 1}, y_j \rangle + \lambda_3 \langle q_{j \oplus k \oplus 1} - q_{j \oplus 1}, y_{j \oplus k} \rangle \\ &> 0. \end{aligned} \qquad \text{(C-17)}$$

Since, by Lemma A-3, $\langle q_{j \oplus k \oplus 1} - q_{j \oplus 1}, y_{j \oplus 1} \rangle > 0$, it follows from (C-17) that $\lambda_2 > 0$.

Case 2: k is even. Then k - 1 is odd and by (C-15)

$$\lambda_2 y_{j \oplus 1} = -\lambda_1 y_j - \lambda_3 y_{j \oplus k}, \tag{C-18}$$

and, using Lemma A-3,

$$\begin{aligned} &\lambda_2 \langle q_{j \oplus k} - q_{j \oplus 1}, y_{j \oplus 1} \rangle \\ &\quad = \lambda_1 \langle q_{j \oplus 1} - q_{j \oplus k}, y_j \rangle + \lambda_3 \langle q_{j \oplus 1} - q_{j \oplus k}, y_{j \oplus k} \rangle \\ &\quad > 0. \end{aligned} \tag{C-19}$$

Since by Lemma A-3, $\langle q_{j \oplus k} - q_{j \oplus 1}, y_{j \oplus 1} \rangle > 0$, it follows from (C-19) that $\lambda_2 > 0$.

It remains to show that $y_j$, $y_{j \oplus 1}$ and $(-1)^k y_{j \oplus k}$ are non-collinear. Since $\lambda_1, \lambda_2, \lambda_3 > 0$, it follows from (C-14) that $0 \in [y_j, y_{j \oplus 1}, (-1)^k y_{j \oplus k}]$. Therefore if $y_j$, $y_{j \oplus 1}$ and $(-1)^k y_{j \oplus k}$ are collinear then they must all lie on a line through the origin. But since we have already shown that $0 \in \text{int}[y_j, y_{j \oplus 1}, y_{j \oplus 2}]$, $y_j$ and $y_{j \oplus 1}$ cannot lie on a line through the origin. Therefore $y_j$, $y_{j \oplus 1}$ and $(-1)^k y_{j \oplus k}$ are non-collinear and hence $0 \in \text{int}[y_j, y_{j \oplus 1}, (-1)^k y_{j \oplus k}]$. This completes that proof of Lemma C-2.

In order to prove that the loop is not infinite it will be shown that the parameter n in the program in Section 3.2.5 can fail to be incremented on at most T - 2 consecutive passes through the loop. The proof will be by contradiction. Accordingly, assume that n is not incremented on T - 1 consecutive passes through the loop.

Let k and $\phi$ have the values that they have at Step 2 of the first of these T - 1 passes. Let

$$b_a = \min\{j:\ 0 \le j \le N - T - 1 \text{ and } \phi(d_k \oplus a,\ j) = 1\} \tag{C-20}$$

for a = 0, . . ., T - 2. Then

$$\phi(d_k \oplus a, b_a) = 1 \tag{C-21}$$

and

$$\phi(a_k \oplus a - d_k \oplus a, b_a) = 1 \tag{C-22}$$

for a = 0, . . ., T - 2. Let

$$x_a = \begin{cases} d_k \oplus a, b_a & \text{if } h_k \oplus a(d_k \oplus a, b_a) \ge 0 \\ a_k \oplus a - d_k \oplus a, b_a & \text{otherwise.} \end{cases} \tag{C-23}$$

Then

$$h_k \oplus a(x_a) = |h_k \oplus a(d_k \oplus a, b_a)| \quad . \tag{C-24}$$

If $x \in Z^2$ and $\phi(x) = 1$, then $x \in D$ and $x = d_k \oplus a,\ j$ for some $j \ge b_a$. Therefore $|h_k \oplus a(x)| = |h_k \oplus a(d_k \oplus a,\ j)| \le |h_k \oplus a\ (d_k \oplus a, b_a)| = h_k \oplus a(x_a)$. Thus, for $x \in Z^2$,

$$\phi(x) = 1 \longrightarrow |h_k \oplus a(x)| \le h_k \oplus a(x_a). \tag{C-25}$$

Claim: For a = 0, . . ., T - 2,

$$(-1)^a \langle x_a - a_k \oplus a \oplus 1,\ y_k \oplus (T-1) \rangle > 0. \tag{C-26}$$

Proof of Claim: First, we will prove the claim for $a = 0$. Since by (C-21), (C-22) and (C-23), $\phi(\alpha_k - x_0) = 1$, it follows that $\alpha_k - x_0 \in D$. Therefore, by Lemma A-3,

$$\langle \alpha_k - x_0, y_k \oplus (T-1) \rangle < \langle q_k, y_k \oplus (T-1) \rangle , \tag{C-27}$$

or, since $\alpha_k = q_k + q_k \oplus 1$,

$$\langle x_0 - q_k \oplus 1, y_k \oplus (T-1) \rangle > 0. \tag{C-28}$$

Thus, the claim is true for $a = 0$. Now let $1 \leq a \leq T - 2$ and assume the claim is true for $a - 1$. By (C-21), (C-22), and (C-23), $\phi(\alpha_k \oplus a - x_a) = 1$ and hence using (C-25),

$$\begin{aligned} h_k \oplus (a-1)(\alpha_k \oplus a - x_a) &\leq |h_k \oplus (a-1)(\alpha_k \oplus a - x_a)| \\ &\leq h_k \oplus (a-1)(x_{a-1}), \end{aligned} \tag{C-29}$$

or equivalently,

$$\langle x_{a-1} - \alpha_k \oplus a + x_a, y_k \oplus (a-1) \rangle \geq 0. \tag{C-30}$$

Similarly, $\phi(x_{a-1}) = 1$ and

$$\begin{aligned} -h_k \oplus a(x_{a-1}) \leq = |h_k \oplus a(x_{a-1})| \\ \leq h_k \oplus a(x_a) \\ = -h_k \oplus a(\alpha_k \oplus a - x_a), \end{aligned} \tag{C-31}$$

and therefore,

$$h_k \oplus a(\alpha_k \oplus a - x_a) \leq h_k \oplus a(x_{a-1}), \tag{C-32}$$

or equivalently,

$$\langle x_{a-1} - \alpha_k \oplus a + x_a,\ y_k \oplus a \rangle \geq 0. \tag{C-33}$$

By Lemma C-2, $0 \in \text{int}[y_k \oplus (a-1),\ y_k \oplus a,\ (-1)^{T-a}\ y_k \oplus (T-1)]$ and therefore there exist strictly positive real numbers $\sigma_1$, $\sigma_2$, $\sigma_3$ such that

$$\sigma_1 y_k \oplus (a-1) + \sigma_2 y_k \oplus a + \sigma_3 (-1)^{T-a}\ y_k \oplus (T-1) = 0. \tag{C-34}$$

Since, by Lemma A-2, T is odd, $(-1)^{T-a} = -(-1)^a$, and from (C-34) we obtain,

$$\sigma_3 (-1)^a\ y_k \oplus (T-1) = \sigma_1 y_k \oplus (a-1) + \sigma_2 y_k \oplus a. \tag{C-35}$$

By (C-35), (C-30) and (C-33),

$$\begin{aligned} &\sigma_3 (-1)^a \langle x_{a-1} - \alpha_k \oplus a + x_a,\ y_k \oplus (T-1) \rangle \\ &= \sigma_1 \langle x_{a-1} - \alpha_k \oplus a + x_a,\ y_k \oplus (a-1) \rangle \\ &\quad + \sigma_2 \langle x_{a-1} - \alpha_k \oplus a + x_a,\ y_k \oplus a \rangle \\ &\quad \geq 0, \end{aligned} \tag{C-36}$$

and since $\sigma_3 > 0$,

$$(-1)^a \langle x_{a-1} - \alpha_k \oplus a + x_a,\ y_k \oplus (T-1) \rangle \geq 0. \tag{C-37}$$

Substituting $\alpha_k \oplus a = q_k \oplus a + q_k \oplus a \oplus 1$ and using (C-37) and the induction hypothesis,

$$
\begin{aligned}
&(-1)^a \langle x_a - q_k \oplus a \oplus 1,\ y_k \oplus (T-1) \rangle \\
&\quad \geq -(-1)^a \langle x_{a-1} - q_k \oplus a,\ y_k \oplus (T-1) \rangle \\
&\quad = (-1)^{a-1} \langle x_{a-1} - q_k \oplus a,\ y_k \oplus (T-1) \rangle \\
&\quad > 0.
\end{aligned}
\tag{C-38}
$$

This completes the proof of the claim.

Now set $a = T - 2$ in (C-26). Since, by Lemma A-2, $T - 2$ is odd we obtain,

$$
\langle x_{T-2},\ y_k \oplus (T-1) \rangle < \langle q_k \oplus (T-1),\ y_k \oplus (T-1) \rangle \ . \tag{C-39}
$$

It now follows from Lemma A-3 that $x_{T-2} \notin D$ and therefore $\phi(x_{T-2}) = 0$ which contradicts either (C-21) or (C-22). Therefore n cannot fail to be incremented on each of $T - 1$ consecutive passes through the loop. This completes the proof that the loop is not infinite.

It now follows that the program produces sequences $q_T, \ldots, q_{N-1}$ and $m_T, \ldots, m_{N-1}$. It remains to show that if $q = (q_0, \ldots, q_{N-1})$ and $m = (m_T, \ldots, m_{N-1})$ then $(q, m)$ is a reconstruction algorithm for the mask M.

We have $R(M) = \{ q_0, \ldots, q_{T-1} \}$. Let $T \leq n \leq N - 1$ and let all variables have the values that they have after Step 5 and before Step 6 of the pass through the loop in which $q_n$ and $m_n$ are defined. By the definition of b in Step 1 of the loop, for $x \in Z^2$,

$$|h_k(d_{k,b})| < |h_k(x)| \longrightarrow \phi(x) = 0. \qquad \text{(C-40)}$$

If $x \in D$, then before entering the loop $\phi$ had to have the value 1 at $x$. Thus if the current value is $\phi(x) = 0$, $\phi$ must have acquired the value 0 at $x$ on some preceeding pass. That is, $x = q_{n'}$ for some $n' < n$. Therefore $x \in \{q_0, \ldots, q_{n-1}\}$. Thus, for $x \in D$,

$$\phi(x) = 0 \longrightarrow x \in \{q_0, \ldots, q_{n-1}\}. \qquad \text{(C-41)}$$

First, it will be shown that

$$M \cap (M + q_n - q_{m_n}) \subseteq \{q_0, \ldots, q_n\}. \qquad \text{(C-42)}$$

Let $x \in M \cap (M + q_n - q_{m_n})$. We want to show that $x \in \{q_0, \ldots, q_n\}$. If $x \in R(M)$ then, since $n \geq T$, we are done. Since $q_n = d_{k,b}$. if $x = d_{k,b}$ we are done. Now assume $x \notin R(M)$ and $x \neq d_{k,b}$. Since $x \notin R(M)$, $x \in D$.

Claim:

$$|h_k(d_{k,b})| < |h_k(x)| \ . \qquad \text{(C-43)}$$

Proof of Claim: The proof of the claim will be divided into two cases.

Case 1:

$$h_k(d_{k,b}) \geq 0. \qquad \text{(C-44)}$$

Since $q_n = d_{k,b}$, it follows from Step 5 of the loop that $m_n = k$. Let $z = x - d_{k,b} + q_k$. Since $x \in M + d_{k,b} - q_k$, it follows that $z \in M$. Also, since $x \neq d_{k,b}$, $z \neq q_k$. Therefore, by Lemma A-3,

$$\langle q_k, y_k \rangle < \langle z, y_k \rangle$$

$$= \langle x, y_k \rangle - \langle d_{k,b}, y_k \rangle + \langle q_k, y_k \rangle , \quad \text{(C-45)}$$

or $\langle d_{k,b}, y_k \rangle < \langle x, y_k \rangle$. It follows that $h_k(d_{k,b}) < h_k(x)$ and since $h_k(d_{k,b}) \geq 0$, the claim follows.

Case 2:

$$h_k(d_{k,b}) < 0. \quad \text{(C-46)}$$

In this case $m_n = k \oplus 1$. Let $z = x - d_{k,b} + q_{k \oplus 1}$. Since $x \in M + d_{k,b} - q_{k \oplus 1}$, it follows that $z \in M$. Also, since $x \neq d_{k,b}$, $z \neq q_{k \oplus 1}$. Therefore, by Lemma A-3,

$$\langle x, y_k \rangle - \langle d_{k,b}, y_k \rangle + \langle q_{k \oplus 1}, y_k \rangle$$

$$= \langle z, y_k \rangle$$

$$< \langle q_{k \oplus 1}, y_k \rangle , \quad \text{(C-47)}$$

or $\langle x, y_k \rangle < \langle d_{k,b}, y_k \rangle$. It follows that $h_k(x) < h_k(d_{k,b})$ and since $h_k(d_{k,b}) < 0$, the claim follows. This completes the proof of the claim.

Now it follows by implication (C-40) that $\phi(x) = 0$ and since $x \in D$, it follows from implication (C-41) that $x \in \{q_0, \ldots, q_{n-1}\} \subseteq \{q_0, \ldots, q_n\}$. This completes the proof of (C-42).

It remains to show that

$$M \cap (M - d_{k,b} + q_{m_n}) \subseteq \{q_0, \ldots, q_{n-1}\} . \quad \text{(C-48)}$$

Let $x \in M \cap (M - d_{k,b} + q_{m_n})$. We want to show that $x \in \{q_0, \ldots, q_{n-1}\}$. Since $n \geq T$, if $x \in R(M)$ we are done. Now assume $x \notin R(M)$. Then $x \in D$. The proof will be divided into two cases.

Case 1: $h_k(d_{k,b}) \geq 0$.

In this case $m_n = k$. Let $z = x + d_{k,b} - q_k$. Since $x \in M - d_{k,b} + q_k$, it follows that $z \in M$.

If $z = q_k \oplus 1$, then $x = q_k + q_k \oplus 1 - d_{k,b} = \alpha_k - d_{k,b}$. Now if $\phi(x) = 1$, then by Step 2 of the loop, $q_n$ would not have been defined on this pass contrary to the assumption that it was. Therefore, $\phi(x) = 0$ and by implication (C-41), $x \in \{q_0, \ldots, q_{n-1}\}$.

Now assume $z \neq q_k \oplus 1$. Then by Lemma A-3,

$$\begin{aligned} &\langle x, y_k \rangle + \langle d_{k,b}, y_k \rangle - \langle q_k, y_k \rangle \\ &\quad = \langle z, y_k \rangle \\ &\quad < \langle q_k \oplus 1, y_k \rangle . \end{aligned} \tag{C-49}$$

Recalling that $\beta_k = (q_k + q_k \oplus 1)/2$, it follows from (C-49) that

$$\langle d_{k,b} - \beta_k, y_k \rangle < \langle -x + \beta_k, y_k \rangle \tag{C-50}$$

or $h_k(d_{k,b}) < -h_k(x)$ and since $h_k(d_{k,b}) \geq 0$, it follows that $|h_k(d_{k,b})| < |h_k(x)|$. Now by implication (C-40), $\phi(x) = 0$ and by implication (C-41), $x \in \{q_0, \ldots, q_{n-1}\}$.

Case 2: $h_k(d_{k,b}) < 0$.

In this case $m_n = k \oplus 1$. Let $z = x + d_{k,b} - q_{k \oplus 1}$. Since $x \in M - d_{k,b} + q_{k \oplus 1}$, it follows that $z \in M$.

If $z = q_k$ then $x = q_k + q_{k \oplus 1} - d_{k,b} = \alpha_k - d_{k,b}$ and by the argument given in Case 1 above, $\phi(x) = 0$, and by implication (C-41), $x \in \{q_0, \ldots, q_{n-1}\}$.

Now assume $z \neq q_k$. Then by Lemma A-3,

$$\langle q_k, y_k \rangle < \langle z, y_k \rangle$$

$$= \langle x, y_k \rangle + \langle d_{k,b}, y_k \rangle - \langle q_{k \oplus 1}, y_k \rangle \quad \text{(C-51)}$$

from which it follows that $-h_k(x) < h_k(d_{k,b})$. Since $h_k(d_{k,b}) < 0$, it follows that $|h_k(d_{k,b})| < |h_k(x)|$. Hence by implication (C-40), $\phi(x) = 0$ and by implication (C-41), $x \in \{q_0, \ldots, q_{n-1}\}$. This establishes (C-48) and completes the proof that (q, m) is a reconstruction algorithm for the mask M.

# APPENDIX D

# PARAMETER ESTIMATION AND THE CRAMER-RAO LOWER BOUND

## D.1. INTRODUCTION

The estimation problem we consider is to estimate a parameter vector $\underline{a} = (a_1, \ldots, a_L)$ from a measurement vector $\underline{R} = (R_1, \ldots, R_M)$. The estimate of $\underline{a}$ is denoted $\hat{\underline{A}} = \hat{\underline{A}}(\underline{R})$. It is assumed that the conditional probability density $p(\underline{r} \mid \underline{a})$ is known. We consider two estimation environments that differ primarily in their prior knowledge of $\underline{a}$.

A word on notation is in order. In general, we will use upper case letters to denote random variables or vectors and will use lower case letters to denote their possible values and also most non-random variables or vectors. Vectors will be underscored. Thus, the received vector, which contains randomness, is denoted $\underline{R}$. A function of a random quantity is also to be considered random; hence, the estimate $\hat{\underline{A}}$ is capitalized. As described below, the parameter vector to be estimated may be random or not. In the previous paragraph, we elected to use the lower case $\underline{a}$. The probability density of a random vector such as $\underline{R}$ will be denoted $p_{\underline{R}}(\underline{r})$, or simply $p(\underline{r})$, when no confusion can arise. A conditional probability density such as that of $\underline{R}$ given $\underline{A}$ will be denoted $p_{\underline{R} \mid \underline{A}}(\underline{r} \mid \underline{a})$ or $p(\underline{r} \mid \underline{a})$. The average of a random quantity will be denoted either $E[A]$ or $E_A[a]$. The latter is especially helpful for expressions like $E_A[dp_A(a)/da]$, which otherwise would have to be written $E[dp_A(A)/dA]$

**Environment 1: Nonrandom Parameters**

Here, the parameter vector to be estimated is a fixed but unknown vector $\underline{a}$, and the quality of an estimator is judged by the *mean squared errors* (MSE)

$$E[(\hat{A}_i - a_i)^2], \quad i = 1, 2, \ldots, L \quad . \tag{D-1}$$

There may be no certifiably best estimator. The *maximum likelihood* (ML) estimator chooses $\hat{A}(\underline{r}) = \underline{a}$ if $p(\underline{r} \mid \underline{a})$ is largest among all choices for $\underline{a}$. An estimator is *unbiased* if $E[\hat{A}_i] = a_i$, $i = 1, \ldots, L$. The ML estimator might or might not be unbiased.

**Environment 2: Random Parameters**

Here, the parameter vector to be estimated is a random vector $\underline{A}$ with known probability density $p(\underline{a})$. The quality of an estimator is judged by the mean squared errors (MSE)

$$E[(\hat{A}_i - A_i)^2], \quad i = 1, \ldots, L \quad . \tag{D-2}$$

The best estimator, i.e., the *minimum mean squared error* (MMSE) estimator is

$$\hat{A}(\underline{r}) = E[\underline{A} \mid \underline{R} = \underline{r}] \tag{D-3}$$

An estimator is *unbiased* if $E[\hat{A}_i] = E[A_i]$, $i = 1, \ldots, L$. The MMSE estimator is unbiased. In either environment, there is a Cramer-Rao lower bound (CRLB) to the MSE.

**Environment 1:** For any unbiased estimator,

$$E[(\hat{A}_i - a_i)^2] \geq [J]_{ii}^{-1}, \quad i = 1, \ldots, L \tag{D-4}$$

$$J_{ij} = E_{\underline{R}}\left\{ \frac{\partial}{\partial a_i} \ln p(\underline{r} \mid \underline{a}) \frac{\partial}{\partial a_j} \ln p(\underline{r} \mid \underline{a}) \right\} \tag{D-5a}$$

$$= - E_{\underline{R}}\left\{ \frac{\partial^2}{\partial a_i \partial a_j} \ln p(\underline{r} \mid \underline{a}) \right\} \quad , \tag{D-5b}$$

where the expectations average over $\underline{R}$.

**Environment 2:** For estimators having a *certain "unbiased-like"* property,

$$E[(\hat{A}_i - A_i)^2] \geq [\bar{J} + K]^{-1}_{ii} \quad , \quad i = 1, \ldots, L \quad , \tag{D-6}$$

where

$$\bar{J}_{ij} = E_{\underline{A}}[J_{ij}] \tag{D-7}$$

$$K_{ij} = E_{\underline{A}}\left\{ \frac{\partial}{\partial a_i} \ln p(\underline{a}) \frac{\partial}{\partial a_j} \ln p(\underline{a}) \right\} \tag{D-8a}$$

$$= - E_{\underline{A}}\left\{ \frac{\partial^2}{\partial a_i \partial a_j} \ln p(\underline{a}) \right\} \tag{D-8b}$$

In this environment, the expectations average over both $\underline{R}$ and $\underline{A}$. In the scalar case (i.e., $L = 1$) the "unbiased-like" property which the estimator must satisfy is

$$\lim_{a \to \pm\infty} p_A(a)\left\{ E_{\underline{R}}[\hat{A}(\underline{r}) \mid A = a] - a \right\} = 0 \quad . \tag{D-9}$$

**Notes:**

(1) The existence of the above derivatives is presumed.

(2) The matrices $J$ and $K$ are functions of the likelihood distribution $p(\underline{r} \mid \underline{a})$ and the a priori distribution $p(\underline{a})$, respectively. $\bar{J}$ is related primarily to the likelihood distribution $p(\underline{r} \mid \underline{a})$ and secondarily to the a priori distribution $p(\underline{a})$.

In the case where $A$ is a scalar ($L = 1$), these bounds reduce to the following.

**Environment 1:**

$$E[(\hat{A} - a)^2] \geq \frac{1}{-E_{R}\left(\frac{\partial^2}{\partial a^2} \ln p(\underline{r} \mid a)\right)} = \frac{1}{E_{R}\left(\left(\frac{\partial}{\partial a} \ln p(\underline{r} \mid a)\right)^2\right)} \quad \text{(D-10)}$$

**Environment 2:**

$$E[(\hat{A} - A)^2] \geq \frac{1}{-E_{R,A}\left(\frac{\partial^2}{\partial a^2} \ln p(\underline{r} \mid a)\right) - E_{A}\left(\frac{\partial^2}{\partial a^2} \ln p(a)\right)} \quad \text{(D-11a)}$$

$$= \frac{1}{E_{R,A}\left(\left(\frac{\partial}{\partial a} \ln p(\underline{r} \mid a)\right)^2\right) + E_{A}\left(\left(\frac{\partial}{\partial a} \ln p(a)\right)^2\right)} \quad \text{(D-11b)}$$

Note that in Environment 2 one may use the bounds in Eq. (D-11) for the vector case as well; i.e.,

$$E[(\hat{A}_i - A_i)^2] \geq \left(E_{R,A_i}\left(\left(\frac{\partial}{\partial a_i} \ln p(\underline{r} \mid a_i)\right)^2\right) + E_{A_i}\left(\left(\frac{\partial}{\partial a_i} p(a_i)\right)^2\right)\right)^{-1} \quad \text{(D-12)}$$

This bound is presumably simpler to compute but not as good as that of Eq. (D-6).

## D.2 ADDITIVE NOISE PROBLEMS

In this section, we focus on the following "additive noise" estimation problem and special cases thereof:

$$\underline{R} = f(\underline{A}) + \underline{N} \quad \text{(D-13)}$$

where $f$ is a known function and $\underline{N} = (N_1, \ldots, N_M)^T$ is a random noise vector, independent of $\underline{A}$, with density $p_{\underline{N}}(\underline{n})$. The conditional density of $\underline{R}$ given $\underline{A} = \underline{a}$ is

$$p(\underline{x} \mid \underline{a}) = p_{\underline{N}}(\underline{x} - \underline{f}(\underline{a})) \quad . \tag{D-14}$$

### D.2.1 SCALAR CASE

When both the parameter to be estimated and the measurement are scalars, the CRLB reduces to:

**Environment 1:**

$$E[(\hat{A} - a)^2] \geq \left[ E_N \left( \left( \frac{p'_N(n)}{p_N(n)} \right)^2 \right) (f'(a))^2 \right]^{-1} \tag{D-15}$$

where $p'_N(n)$ denotes the derivative of $p_N(n)$ with respect to $n$, and $f'(a)$ denotes the derivative of $f(a)$ with respect to $a$.

**Environment 2:**

$$E[(\hat{A} - A)^2] \geq \left[ E_N \left( \left( \frac{p'_N(n)}{p_N(n)} \right)^2 \right) E[(f'(A))^2] + E_A \left( \left( \frac{p'(a)}{p(a)} \right)^2 \right) \right]^{-1}, \tag{D-16}$$

where $p'(a)$ denotes the derivative of $p(a)$ with respect to $a$. Note that the functional form of $f$ influences only the first term in the brackets, whereas the a priori distribution influences both terms.

### D.2.2 VECTOR CASE

One may obtain similar but more complicated expressions for the CRLB for the vector case. We note that the functional form of $f$ influences $J$ only, whereas the a priori distribution influences both $\bar{J}$ and $K$.

### D.2.3 IID GAUSSIAN CASE

Here $\underline{N} = (N_1, \ldots, N_M)$ is a vector of independent and identically distributed (IID) Gaussian $\eta(0,\sigma_N^2)$ random variables, and (in environment 2) $\underline{A} = (A_1, \ldots, A_L)$ is IID Gaussian, $\eta(0,\sigma_A^2)$. The CRLB becomes:

**Environment 1:**

$$E[(\hat{A}_i - a_i)^2] \geq [J]_{ii}^{-1} \quad , \tag{D-17}$$

$$J_{ij} = \frac{1}{\sigma_N^2} \sum_{k=1}^{M} \frac{\partial}{\partial a_i} f_k(\underline{a}) \frac{\partial}{\partial a_j} f_k(\underline{a}) \tag{D-18a}$$

$$\left( J = \frac{1}{\sigma_N^2} (\nabla_a f)^T (\nabla_a f) \right) , \left( [\nabla_a f]_{ki} \triangleq \frac{\partial}{\partial a_i} f_k \right) \tag{D-18b}$$

**Environment 2:**

$$E[(\hat{A}_i - A_i)^2] \geq [\bar{J} + K]_{ii}^{-1} \quad , \tag{D-19}$$

$$\bar{J}_{ij} = E_{\underline{A}}[J_{ij}] = \frac{1}{\sigma_N^2} \sum_{k=1}^{M} E_{\underline{A}} \left( \frac{\partial}{\partial a_i} f_k(\underline{a}) \frac{\partial}{\partial a_j} f_k(\underline{a}) \right) \tag{D-20}$$

$$K_{ij} = \frac{1}{\sigma_A^2} \delta_{ij} \quad , \quad (\delta_{ij} = 0 \, , \; i \neq j \text{ and } \delta_{ii} = 1) \tag{D-21a}$$

$$(K = \frac{1}{\sigma_A^2} I \, , \; I = \text{identity matrix}) \tag{D-21b}$$

In the scalar case with $L = M = 1$, these reduce to

**Environment 1:**

$$E[(\hat{A} - a)^2] \geq \frac{\sigma_N^2}{(f'(a))^2} \tag{D-22}$$

**Environment 2:**

$$E[(\hat{A} - A)^2] \geq \frac{1}{E[(f'(A))^2]/\sigma_N^2 + 1/\sigma_A^2} \tag{D-23}$$

or

$$\frac{E[\hat{A} - A]^2}{\sigma_A^2} \geq \frac{1}{E[(f'(A))^2]\,\sigma_A^2/\sigma_N^2 + 1} \tag{D-24}$$

### D.2.4 LINEAR IID GAUSSIAN CASE

Here

$$\underline{R} = F\underline{A} + \underline{N} \ , \tag{D-25}$$

where $F$ is an $M \times L$ matrix; $\underline{R}$, $\underline{A}$ and $\underline{N}$ are column vectors; and $\underline{A}$, $\underline{N}$ are independent and IID Gaussian as before. In this case, the CRLB becomes:

**Environment 1:**

$$E[(\hat{A}_i - a_i)^2] \geq [J]_{ii}^{-1} = \sigma_N^2 [F^T F]_{ii}^{-1} \tag{D-26}$$

$$J_{ij} = \frac{1}{\sigma_N^2} \sum_{k=1}^{M} F_{ki} F_{kj} = \frac{[F^T F]_{ij}}{\sigma_N^2} \tag{D-27}$$

**Environment 2:**

$$E[(\hat{A}_i - A_i)^2] \geq [\bar{J} + K]_{ii}^{-1} = \left[\frac{F^T F}{\sigma_N^2} + \frac{I}{\sigma_A^2}\right]_{ii}^{-1} \tag{D-28}$$

$$\bar{J} = J - \frac{F^T F}{\sigma_N^2} \tag{D-29}$$

$$K_{ij} = \frac{\delta_{ij}}{\sigma_A^2} \quad , \quad \left( K = \frac{1}{\sigma_A^2} I \right) \tag{D-30}$$

Additionally, it is known that there exist linear estimators for which the bounds are tight, i.e., for which equality holds.

If $F$ is an orthogonal matrix, i.e, its rows are orthonormal, its columns are orthonormal, and $F^{-1} = F^T$, then

**Environment 1:**

$$E[(\hat{A}_i - a_i)^2] \geq \sigma_N^2 \tag{D-31}$$

**Environment 2:**

$$E[(\hat{A}_i - A_i)^2] \geq \frac{\sigma_A^2}{1 + \sigma_A^2 / \sigma_N^2} \tag{D-32}$$

If $F$ is orthogonal except for a scale factor, i.e., its rows are orthogonal, with norm $c$, then $F^T F = c^2 I$, and

**Environment 1:**

$$E[(\hat{A}_i - a_i)^2] \geq \frac{\sigma_N^2}{c^2} \tag{D-33}$$

**Environment 2:**

$$E[(\hat{A}_i - A_i)^2] \geq \frac{\sigma_A^2}{1 + c^2 \sigma_A^2 / \sigma_N^2} \tag{D-34}$$

If $F$ is a... with $F_{ii} = W_i$, as it would be if it represented a simple weighting of the components of $A$, then

**Environment 1:**

$$E[(\hat{A}_i - a_i)^2] \geq \frac{\sigma_N^2}{W_i^2} \tag{D-35}$$

**Environment 2:**

$$E[(\hat{A}_i - A_i)^2] \geq \frac{\sigma_A^2}{1 + W_i^2 \sigma_A^2 / \sigma_N^2} \tag{D-36}$$

## D.3 CRAMER-RAO BOUNDS IN THE PRESENCE OF AMBIGUITY

Consider the additive noise estimation problem where

$$R = f(A) + N \quad . \tag{D-37}$$

If $f$ is not one-to-one, we say that there is *ambiguity* in the function $f$ and the measurement $R$. In such cases, one cannot expect any estimator to perform well. Unfortunely, however, the Cramer-Rao bound may not reflect this, i.e, it may give a very low MSE even though the actual MSE is rather large. We illustrate this phenomenon in the scalar case and show that it is not a problem in the linear Gaussian case.

**Example 1:** Scalar Parameters

One may see from Eq.s (D-15) and (D-16) and, for the Gaussian case, Eq.s (D-22) and (D-23) that the CRLB will be the same for any function $\tilde{f}$ whose derivative has the same magnitude as that of $f$. To get a clearer picture, consider the situation where $f$ has a continuous second derivative but is not one-to-one. Then $f$ has intervals where it is constant ($f' = 0$) and/or pairs of intervals $I_+$, $I_-$ such that $f$ increases on $I_+$ ($f' > 0$) and $f$ decreases through the same range on $I_-$ ($f' < 0$). The CRLB will be made large (and appropriately so) by the intervals where $f' = 0$. On the other hand, it

will not be affected by the existence of pairs of intervals where $f$ increases and decreases, respectively, through the same range, i.e., it will be as small as the CRLB for the nonambiguous function

$$\dot{f}(a) \triangleq \int_{-\infty}^{a} |f'(z)|\, dz \quad , \tag{D-38}$$

for which one expects there are estimators with much lower MSE than for $f$. Thus, we conclude that in the presence of ambiguity, the CRLB can only be expected to give a reasonable lower bound to the MSE for the estimation problem involving $\dot{f}$.

As a concrete example, compare the estimation of $A$ based on either but not both of the following two measurements

$$R_1 = f_1(A) + N \ , \ R_2 = f_2(A) + N \ , \tag{D-39}$$

where $f_1(a) = a^3$, $f_2(a) = |a^3|$, $A$ and $N$ are Gaussian, independent and $\eta(0, \sigma_A^2)$, $\eta(0, \sigma_N^2)$, respectively. There is obvious ambiguity in the $R_2$ measurement. For either measurement, the CRLB gives

$$E[(\hat{A} - A)^2] \geq \frac{1}{27\sigma_A^4/\sigma_N^2 + 1/\sigma_A^2} \tag{D-40}$$

This is a reasonable bound for estimation based on $R_1$. (It tends to 0 as $\sigma_N^2 \to 0$; it tends to $\sigma_A^2$ as $\sigma_N^2 \to \infty$.) It is not reasonable for estimation based on $R_2$. Indeed if $\sigma_N^2 = 0$, then the best estimate for $A$ based on $R_2$ is $\hat{A} = E[A \mid R_2] = 0$ with MSE $= \sigma_A^2$, whereas the CRLB lower bound is zero.

Finally, consider estimation based on $R_2$, when the density of $A$ is replaced by

$$\tilde{p}(a) = \begin{cases} \frac{2}{\sqrt{2\pi}\sigma_A} \exp\{ -\frac{a^2}{2\sigma_A^2} \} , & a \geq 0 \\ 0 , & a < 0 . \end{cases} \tag{D-41}$$

Now the increased a priori information removes all ambiguity (at least with probability 1), and the CRLB (which is the same as before) is a reasonable bound to MSE based on $R_2$. Thus, we see that additional a priori information can modify the problem so the CRLB is useful.

**Example 2:** Linear IID Gaussian Case

Here,

$$\underline{R} = F\underline{A} + \underline{N} , \tag{D-42}$$

where $\underline{R} = (R_1, \ldots, R_M)^T$, $\underline{A} = (A_1, \ldots, A_L)^T$, $\underline{N} = (N_1, \ldots, N_M)^T$, $F$ is an $M \times L$ matrix, and $\underline{A}$ and $\underline{N}$ are IID Gaussian $\eta(0, \sigma_A^2)$ and $\eta(0, \sigma_N^2)$, respectively and independent of each other. For MMSE estimator, the CRLB bound (19) is tight and gives

$$E[(A_i - \hat{A}_i)^2] = \left[ \frac{F^T F}{\sigma_N^2} + \frac{I}{\sigma_A^2} \right]^{-1}_{ii} \tag{D-43}$$

If $F$ has rank less than $L$ (for example, if $M < L$), there will be ambiguity in $\underline{R}$. Nevertheless, as mentioned in Section D.2.4, the CRLB is tight.

As a concrete example, suppose

$$R = A_1 + A_2 + N , \tag{D-44}$$

so that $L = 2$, $M = 1$, $F = [1\ 1]$. Then

$$\bar{J}+K = \frac{1}{\sigma_N^2}\begin{bmatrix}1 & 1\\ 1 & 1\end{bmatrix} + \frac{1}{\sigma_A^2}\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}, \tag{D-45}$$

$$[\bar{J}+K]^{-1} = \frac{\sigma_A^4\,\sigma_N^2}{\sigma_N^2+2\sigma_A^2}\begin{bmatrix}\dfrac{\sigma_N^2+\sigma_A^2}{\sigma_A^2\,\sigma_N^2} & -\dfrac{1}{\sigma_N^2}\\ -\dfrac{1}{\sigma_N^2} & \dfrac{\sigma_N^2+\sigma_A^2}{\sigma_A^2\,\sigma_N^2}\end{bmatrix} \tag{D-46}$$

and for the MMSE estimator it is well known that

$$E[(A_1-\hat{A}_1)^2] = E[(A_2-\hat{A}_2)^2] = \frac{\sigma_A^2(\sigma_N^2+\sigma_A^2)}{\sigma_N^2+2\sigma_A^2} \tag{D-47}$$

Note that as $\sigma_N^2 \to 0$, MSE $\to \sigma_A^2/2$, and as $\sigma_N^2 \to \infty$, MSE $\to \sigma_A^2$.

## Appendix E
## $\nabla e_s^2(g(x))$ FOR COMPLEX OBJECTS

In this appendix we generalize the expression for the gradient of the summed objective function to include objects and object estimates that are complex valued.

Recall that the summed objective function is defined as the sum of a generalized object-domain error metric and the Fourier-domain error metric:

$$e_s^2 = \epsilon_o^2 + e_F^2, \tag{E-1}$$

where

$$\epsilon_o^2 = \sum_{x \epsilon S'} |g(x)|^2, \tag{E-2}$$

and

$$e_F^2 = N^{-2} \sum_u [\,|G(u)| - |F(u)|\,]^2. \tag{E-3}$$

Notice the summed objective function is implicitly a function of the real and imaginary parts of the pixel values for the latest estimate, $g(x)$. We therefore treat the real and imaginary parts of each pixel as distinct parameters that can be adjusted in order to minimize $e_s^2$. We express the real and imaginary parts of the latest estimate as

$$g(x) = a(x) + ib(x). \tag{E-4}$$

We define the gradients to be:

$$\nabla e_s^2(g(x)) = \sum_{j=1}^{N^2} \frac{\partial e_s^2}{\partial a(x_j)} v_j^R + \frac{\partial e_s^2}{\partial b(x_j)} v_j^I \tag{E-5}$$

where $v_j^R$ and $v_j^I$ are orthogonal unit vectors associated with the real and imaginary parts of the $j^{th}$ pixel. The first partial derivative in Eq. (E-5) may be separated into two terms:

$$\frac{\partial e_s^2}{\partial a(x_j)} = \frac{\partial e_F^2}{\partial a(x_j)} + \frac{\partial e_o^2}{\partial a(x_j)} \quad . \tag{E-6}$$

For the moment we examine the first term in (E-6):

$$\frac{\partial e_F^2}{\partial a(x_j)} = \frac{\partial}{\partial a(x_j)} N^{-2} \sum_u \left[ |G(u)| - |F(u)| \right]^2 \tag{E-7}$$

$$= 2N^{-2} \sum_u (|G(u)| - |F(u)|) \frac{\partial |G(u)|}{\partial a(x_j)} \quad . \tag{E-8}$$

It is easy to see that

$$\frac{\partial |G(u)|}{\partial a(x_j)} = \frac{1}{2|G(u)|} \frac{\partial |G(u)|^2}{\partial a(x_j)}$$

$$= \frac{1}{2|G(u)|} \left[ G(u)\, e^{i2\pi u \cdot x_j/N} + C.C. \right] \quad . \tag{E-9}$$

Substituting Eq. (E-9) into Eq. (E-8) yields

$$\frac{\partial e_F^2}{\partial a(x_j)} = N^{-2} \sum_u (|G(u)| - |F(u)|) \frac{(G(u)\, e^{i2\pi u \cdot x_j / N} + C.C.)}{|G(u)|} \qquad \text{(E-10)}$$

$$= (g(x_j) - g'(x_j)) + (g^*(x_j) - g'^*(x_j)) \qquad \text{(E-11)}$$

$$= 2(a(x_j) - a'(x_j)). \qquad \text{(E-12)}$$

This result is consistent with the result quoted for real-valued objects. A parallel derivation gives

$$\frac{\partial e_F^2}{\partial b(x_j)} = 2(b(x_j) - b'(x_j)) \quad . \qquad \text{(E-13)}$$

We now return to the second term in Eq. (E-6):

$$\frac{\partial \epsilon_o^2}{\partial a(x_j)} = \frac{\partial}{\partial a(x_j)} \sum_{x \in S'} |g(x)|^2$$

$$= \frac{\partial}{\partial a(x_j)} \sum_{x \in S'} a^2(x) + b^2(x)$$

$$= \begin{cases} 2a(x_j) \ , & x_j \in S' \\ 0 \ , & x_j \in S \quad . \end{cases} \qquad \text{(E-14)}$$

Similarly,

$$\frac{\partial \epsilon_o^2}{\partial b(x_j)} = \begin{cases} 2b(x_j) \ , & x_j \in S' \\ 0 \ , & x_j \in S \end{cases} \qquad \text{(E-15)}$$

Let

$$S'(x) = \begin{cases} 1, & x \in S' \\ 0, & x \in S \end{cases} \qquad \text{(E-16)}$$

Now Eqs. (E-14) and (E-15) may be expressed more conveniently:

$$\frac{\partial \varepsilon_o^2}{\partial a(x_j)} = 2a(x_j)S'(x_j) \qquad \text{(E-17)}$$

and

$$\frac{\partial \varepsilon_o^2}{\partial b(x_j)} = 2b(x_j)S'(x_j) \quad . \qquad \text{(E-18)}$$

Collecting these results, we have:

$$\frac{\partial e_s^2}{\partial a(x_j)} = 2[a(x_j) - a'(x_j)] + 2a(x_j)S'(x_j) \qquad \text{(E-19)}$$

$$\frac{\partial e_s^2}{\partial b(x_j)} = 2[b(x_j) - b'(x_j)] + 2b(x_j)S'(x_j) \qquad \text{(E-20)}$$

Equations (E-19) and (E-20) may be combined to form a complex gradient image.

$$\text{Gradient image} \equiv 2[g(x) - g'(x)] + 2g(x)S'(x) \qquad \text{(E-21)}$$

The extension to complex-valued objects still requires only 2 FFTs to compute the gradient.

## Appendix F
## $\nabla e_o^2(g(x))$ FOR REAL OBJECTS

Recall that the object-domain error metric is implicitly a function of the input estimate g(x) and is given by:

$$e_o^2(g(x)) = \sum_{x \in S'} [g'(x)]^2. \qquad \text{(F-1)}$$

Its gradient with respect to the input pixel values may be written:

$$\nabla e_o^2(g(x)) = \sum_{j=1}^{N^2} \frac{\partial e_o^2}{\partial g(x_j)} \, v_j \qquad \text{(F-2)}$$

where $v_j$ is a unit vector in the direction of the parameter $g(x_j)$ in parameter space. We focus now on the partial derivative that appears in Eq. (F-2):

$$\frac{\partial e_o^2}{\partial g(x_j)} = \frac{\partial}{\partial g(x_j)} \sum_{x \in S'} \left[ g'(x) \right]^2$$

$$= 2 \sum_{x \in S'} g'(x) \frac{\partial}{\partial g(x_j)} g'(x) \, . \qquad \text{(F-3)}$$

Substitution of the Fourier-domain expression for g'(x) gives:

$$\frac{\partial e_o^2}{\partial g(x_j)} = 2 \sum_{x \in S'} g'(x) \frac{\partial}{\partial g(x_j)} \left\{ N^{-2} \sum_{u} G'(u) \, e^{i 2\pi u \cdot x / N} \right\} \, . \qquad \text{(F-4)}$$

Recall that

$$G'(u) = \frac{G(u)|F(u)|}{|G(u)|} \qquad \text{(F-5)}$$

giving

$$\frac{\partial e_o^{\,2}}{\partial g(x_j)} = 2 \sum_{x \epsilon S'} g'(x) \frac{\partial}{\partial g(x_j)} \left\{ N^{-2} \sum_u \frac{G(u)|F(u)|}{|G(u)|} e^{i2\pi u \cdot x/N} \right\}$$

$$= 2N^{-2} \sum_{x \epsilon S'} g'(x) \sum_u |F(u)| e^{i2\pi u \cdot x/N} \frac{\partial}{\partial g(x_j)} \left\{ \frac{G(u)}{|G(u)|} \right\} . \qquad (F\text{-}6)$$

In order to evaluate the partial derivative in Eq. (F-6) we need expressions for $\frac{\partial G(u)}{\partial g(x_j)}$ and $\frac{\partial |G(u)|}{\partial g(x_j)}$ :

$$\frac{\partial G(u)}{\partial g(x_j)} = \frac{\partial}{\partial g(x_j)} \sum_x g(x) e^{-i2\pi u \cdot x/N}$$

$$= \sum_x e^{-i2\pi u \cdot x/N} \frac{\partial g(x)}{\partial g(x_j)}$$

$$= e^{-i2\pi u \cdot x_j/N} . \qquad (F\text{-}7)$$

The derivation for $\frac{\partial |G(u)|}{\partial g(x_j)}$ requires the result in Eq. (F-7):

$$\frac{\partial |G(u)|}{\partial g(x_j)} = \frac{1}{2|G(u)|} \frac{\partial}{\partial g(x_j)} |G(u)|^2$$

$$= \frac{1}{2|G(u)|} \left[ G^*(u) \frac{\partial G(u)}{\partial g(x_j)} + C.C. \right]$$

$$= \frac{1}{2|G(u)|} \left[ G^*(u) e^{-i2\pi u \cdot x_j/N} + C.C. \right] \qquad (F\text{-}8)$$

where C.C. stands for complex conjugate of the explicit term. Using the results in (F-7) and (F-8) and with some algebraic manipulation the partial derivative in (F-6) becomes

$$\frac{\partial}{\partial g(x_j)} \frac{G(u)}{|G(u)|} = \frac{G^*(u)\, e^{-i2\pi u \cdot x_j/N}}{2|G(u)|\, G^*(u)} - \text{C.C.} \quad . \tag{F-9}$$

Substituting Eq. (F-9) back into Eq. (F-6) yields

$$\frac{\partial e_o^2}{\partial g(x_j)} = N^{-2} \sum_{x \in S'} g'(x) \sum_u |F(u)| \left[ \frac{G^*(u)\, e^{-i2\pi u \cdot x_j/N}}{|G(u)| G^*(u)} - \text{C.C.} \right] e^{i2\pi u \cdot x/N} \quad . \tag{F-10}$$

By changing the order of summation we have

$$\frac{\partial e_o^2}{\partial g(x_j)} = N^{-2} \sum_u |F(u)| \left[ \frac{G^*(u)\, e^{-i2\pi u \cdot x_j/N}}{|G(u)| G^*(u)} - \text{C.C.} \right] \sum_{x \in S'} g'(x)\, e^{i2\pi u \cdot x/N} \quad . \tag{F-11}$$

At this point it is convenient to define the characteristic function of the complement of the support S as follows

$$S'(x) = \begin{cases} 1 \ , \ x \in S' \\ \\ 0 \ , \ x \in S \end{cases} \quad . \tag{F-12}$$

The second summation in (F-11) may now be rewritten

$$\sum_{x \in S'} g'(x)\, e^{i2\pi u \cdot x/N} = \sum_x S'(x) g'(x)\, e^{i2\pi u \cdot x/N} \quad . \tag{F-13}$$

The error in the output, $g_e(x)$, consists of that component of the output that violates the support constraint:

$$g_e(x) = S'(x) g'(x) \quad . \tag{F-14}$$

The summation in Eq. (F-13) has the form of a forward DFT:

$$\sum_x g_e(x)\, e^{i2\pi u\cdot x/N} = \sum_x g_e(x)\, e^{-i2\pi(-u\cdot x)/N}$$

$$= G_e(-u) \qquad \text{(F-15)}$$

where $G_e(u)$ is the DFT of $g_e(x)$. The total partial derivative may now be written:

$$\frac{\partial e_o^2}{\partial g(x_j)} = N^{-2} \sum_u \frac{|F(u)|G_e(-u)}{|G(u)|G^*(u)}\left[G^*(u)\, e^{-i2\pi u\cdot x_j/N} - C.C.\right]$$

$$= N^{-2} \sum_u \frac{|F(u)|G_e(-u)}{|G(u)|}\, e^{-i2\pi u\cdot x_j/N} - N^{-2} \sum_u \frac{|F(u)|G_e(-u)G(u)}{|G(u)|G^*(u)}\, e^{i2\pi u\cdot x_j/N} \qquad \text{(F-16)}$$

Using Eq. (F-5) in the second term:

$$\frac{\partial e_o^2}{\partial g(x_j)} = N^{-2} \sum_u \frac{|F(u)|G_e(-u)}{|G(u)|}\, e^{-i2\pi u\cdot x_j/N} - N^{-2} \sum_u \frac{G'(u)G_e(-u)}{G^*(u)}\, e^{i2\pi u\cdot x_j/N} \qquad \text{(F-17)}$$

Remarkably both of these terms have the general form of DFTs. In order to combine both of these terms into a single inverse DFT we perform a sign change of variable on the Fourier vector in the first term. The net result is:

$$\frac{\partial e_o^2}{\partial g(x_j)} = N^{-2} \sum_u \left[\frac{|F(-u)|G_e(u)}{|G(-u)|} - \frac{G'(u)G_e(-u)}{G^*(j)}\right] e^{i2\pi u\cdot x_j/N} \qquad \text{(F-18)}$$

Finally, by appealing to the Hermitian property for Fourier transforms of real functions we may make the following substitutions:

$$|F(-u)| = |F(u)|$$

$$|G(-u)| = |G(u)|$$

$$G_e(-u) = G_e^*(u) \qquad \text{(F-19)}$$

to get the final result:

$$\frac{\partial e_o^2}{\partial g(x_j)} = N^{-2} \sum_u \left[ \frac{|F(u)| G_e(u)}{|G(u)|} - \frac{G'(u) G_e^*(u)}{G^*(u)} \right] e^{i2\pi u \cdot x_j / N} . \qquad \text{(F-20)}$$

## Appendix G
## $\nabla e_o^2(g(x))$ FOR COMPLEX OBJECTS

The derivation of the gradient of the $e_o^2(g(x))$ objective function for the case of complex objects closely follows that for real objects (Appendix F). Therefore, we just highlight this derivation, calling attention to significant differences. We begin by acknowledging that the complex case admits twice as many parameters; namely the real and imaginary parts of each input pixel. We denote the real and imaginary parts of the input function as follows:

$$g(x) = a(x) + i\, b(x). \tag{G-1}$$

We define the gradient to be:

$$\nabla e_o^2(g(x)) = \sum_{j=1}^{N^2} \frac{\partial e_o^2}{\partial a(x_j)} v_j^R + \frac{\partial e_o^2}{\partial b(x_j)} v_j^I \tag{G-2}$$

where $v_j^R$ and $v_j^I$ are orthogonal unit vectors associated with the parameters of the real and imaginary parts of the pixel at location $x_j$. The partial derivative of the objective function with respect to the real part of a pixel value may be written

$$\begin{aligned}
\frac{\partial e_o^2}{\partial a(x_j)} &= \frac{\partial}{\partial a(x_j)} \sum_{x \in S'} |g'(x)|^2 \\
&= \sum_{x \in S'} g'^*(x) \frac{\partial g'(x)}{\partial a(x_j)} + C.C. \\
&= \left[ \sum_{x \in S'} g'^*(x) \frac{\partial}{\partial a(x_j)} \left\{ N^{-2} \sum_u G'(u)\, e^{i2\pi u \cdot x/N} \right\} \right] + C.C. \\
&= \left[ N^{-2} \sum_{x \in S'} g'^*(x) \sum_u |F(u)|\, e^{i2\pi u \cdot x/N} \frac{\partial}{\partial a(x_j)} \left\{ \frac{G(u)}{|G(u)|} \right\} \right] + C.C.
\end{aligned} \tag{G-3}$$

The extra complex-conjugate term appears because the output function, g'(x), can assume complex values. The partial derivative with respect to the imaginary part is similarly found:

$$\frac{\partial e_o^2}{\partial b(x_j)} = \left[ N^{-2} \sum_{x \epsilon S'} g'^*(x) \sum_u |F(u)|\, e^{i2\pi u \cdot x/N} \frac{\partial}{\partial b(x_j)} \left\{ \frac{G(u)}{|G(u)|} \right\} \right] + C.C. \quad (G\text{-}4)$$

With simple algebraic manipulations the following useful identities may be verified:

$$\frac{\partial G(u)}{\partial a(x_j)} = e^{-i2\pi u \cdot x_j/N} \quad (G\text{-}5)$$

$$\frac{\partial G(u)}{\partial b(x_j)} = ie^{-i2\pi u \cdot x_j/N} \quad (G\text{-}6)$$

$$\frac{\partial |G(u)|}{\partial a(x_j)} = \frac{1}{2|G(u)|} \left[ G^*(\cdot\cdot)\, e^{-i2\pi u \cdot x_j/N} + C.C. \right] \quad (G\text{-}7)$$

$$\frac{\partial |G(u)|}{\partial b(x_j)} = \frac{i}{2|G(u)|} \left[ G^*(u)\, e^{-i2\pi u \cdot x_j/N} - C.C. \right] . \quad (G\text{-}8)$$

With the aid of Eqs. (G-5) thru (G-8) we deduce

$$\frac{\partial}{\partial a(x_j)} \left\{ \frac{G(u)}{|G(u)|} \right\} = \frac{G^*(u)\, e^{-i2\pi u \cdot x_j/N} - C.C.}{2|G(u)|G^*(u)} \quad (G\text{-}9)$$

and

$$\frac{\partial}{\partial b(x_j)}\left\{\frac{G(u)}{|G(u)|}\right\} = \frac{iG^*(u)\, e^{-i2\pi u\cdot x_j/N} - C.C}{2|G(u)|G^*(u)} \quad . \tag{G-10}$$

When Eq. (G-9) is substituted into Eq. (G-3) and the order of summation is interchanged we get:

$$\frac{\partial e_o^2}{\partial a(x_j)} = \left[\frac{N^{-2}}{2}\sum_u |F(u)|\left\{\frac{G^*(u)\, e^{-i2\pi u\cdot x_j/N}}{|G(u)|G^*(u)} - C.C.\right\}\sum_x S'(x)g'^*(x)\, e^{i2\pi u\cdot x/N}\right] + C.C. \tag{G-11}$$

were S'(x) is the characteristic function of the set S', as before. Recall the object and Fourier-domain expressions for the error image:

$$g_e(x) = S'(x)g'(x) \tag{G-12}$$

$$G_e(u) = \sum_x S'(x)g'(x)e^{-i2\pi u\cdot x/N} \quad . \tag{G-13}$$

We may therefore write

$$G_e^*(u) = \sum_x S'(x)g'^*(x)e^{i2\pi u\cdot x/N} \tag{G-14}$$

which may be substituted into (G-11) to get:

$$\frac{\partial e_o^2}{\partial a(x_j)} = \left[\frac{N^{-2}}{2}\sum_u \frac{|F(u)|G_e^*(u)\left\{G^*(u)\, e^{-i2\pi u\cdot x_j/N} - C.C.\right\}}{|G(u)|G^*(u)}\right] + C.C. \tag{G-15}$$

By similar steps we find

$$\frac{\partial e_o^2}{\partial b(x_j)} = \left[ \frac{N^{-2}}{2} \sum_u \frac{|F(u)|G_e^*(u)\left\{iG^*(u)\, e^{-i2\pi u \cdot x_j/N} - C.C.\right\}}{|G(u)|G^*(u)} \right] + C.C. \quad \text{(G-16)}$$

Explicitly writing out the complex-conjugate terms in Eqs. (G-15) and (G-16) and performing a few additional manipulations we produce the final result.

$$\frac{\partial e_o^2}{\partial a(x_j)} = \mathrm{Re}\left\{ N^{-2} \sum_u \left[ \frac{|F(-u)|G_e^*(-u)}{|G(-u)|} - \frac{G'(u)G_e^*(u)}{G^*(u)} \right] e^{i2\pi u \cdot x_j/N} \right\} \quad \text{(G-17)}$$

$$\frac{\partial e_o^2}{\partial b(x_j)} = -\mathrm{Im}\left\{ N^{-2} \sum_u \left[ \frac{F(-u)G_e^*(-u)}{|G(-u)|} + \frac{G'(u)G_e^*(u)}{G^*(u)} \right] e^{i2\pi u \cdot x_j/N} \right\} \quad \text{(G-18)}$$

It is gratifying that Eq. (G-17) is consistent with Eq. (F-18) which is the equivalent partial derivative for real objects only. It is worth mentioning that because the summation arguments in Eqs. (G-17) and (G-18) differ, an additional FFT is required in the computation of the gradient for complex objects. The total number of FFTs (forward or inverse) needed is increased to five.